FACULTY OF BIOSCIENCE ENGINEERING

Data-driven models and trait-oriented experiments of aquatic macrophytes to support freshwater management

Wout Van Echelpoel

Promoter: Prof. Dr. Peter Goethals

Thesis submitted in fulfilment of the requirements for the degree of Doctor (PhD) in Bioscience Engineering

Academic year: 2019 - 2020

GHENT UNIVERSITY

Members of the jury

Prof. dr. Wim Verbeke (Chairman)

Department of Agricultural Economics
Faculty of Bioscience Engineering
Ghent University

Prof. dr. Diederik Rousseau

Department of Green Chemistry and Technology
Faculty of Bioscience Engineering
Ghent University

Prof. dr. ir. Piet Verdonschot

Freshwater Ecosystems
Wageningen University and Research (WUR)

Prof. dr. Stijn Luca

Department of Data Analysis and Mathematical Modelling
Faculty of Bioscience Engineering
Ghent University

Dr. Dries Landuyt

Department of Environment
Faculty of Bioscience Engineering
Ghent University

Dr. Peter Van Puijenbroek

Department of Water, Agriculture and Food
PBL Netherlands Environmental Assessment Agency

Promoter

Prof. dr. Peter Goethals

Aquatic Ecology Research Unit
Department of Animal Sciences and Aquatic Ecology
Faculty of Bioscience Engineering
Ghent University

Dean

Prof. dr. Marc Van Meirvenne

Department of Environment
Faculty of Bioscience Engineering
Ghent University

Rector

Prof. dr. Rik Van de Walle

Wout Van Echelpoel

Data-driven models and trait-oriented experiments of aquatic macrophytes to support freshwater management

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (PhD) in Bioscience Engineering
Academic year 2019-2020

Dutch translation of the title:

Data-gedreven modellen en eigenschapsgeoriënteerde experimenten van aquatische macrofyten ter ondersteuning van zoetwaterbeheer

Please refer to this work as follows:

Wout Van Echelpoel (2020) Data-driven models and trait-oriented experiments of aquatic macrophytes to support freshwater management. PhD Thesis. Department of Animal Sciences and Aquatic Ecology, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium.

ISBN 978-94-6357-314-6

Preface

Het zijn speciale tijden op het moment dat dit voorwoord geschreven wordt. België zit sinds 14 maart 2020 in een "zachte lockdown" en enkel essentiële verplaatsingen en fysieke vergaderingen zijn toegelaten. Waar iedereen enkele weken voordien nog luchtig over deed, stond dit plots voor de deur met nagenoeg onverwachte gevolgen voor de volledige samenleving. Iets gelijkaardigs vindt plaats tijdens het werken aan en schrijven van een doctoraat: het merendeel van de tijd denk je dat het zo'n vaart niet zal lopen, tot je plots aan je laatste maanden begint. Uiteindelijk ben ik er geraakt, na zes jaar gevuld met professionele (en persoonlijke) uitdagingen die het hele proces alleen maar interessanter hebben gemaakt. Veel van deze uitdagingen brachten nieuwe contacten met zich mee, die (ieder op zijn of haar specifieke manier) hebben bijgedragen tot de finale versie van dit werk. Een woordje van dank is dus zeker gepast.

Allereerst is er natuurlijk mijn promotor, professor Peter Goethals, die mij een mooi assortiment aan kansen voor persoonlijke en professionele ontwikkeling heeft aangeboden. De belangrijkste blijft ontegensprekelijk de uitnodiging om te solliciteren voor de job als assistent die mij tijdens mijn master-na-master jaar werd toegezonden. Ook de vele opties om deel te nemen aan buitenlandse campagnes (o.a. Ecuador, Oeganda, Ethiopië en Vietnam) en persoonlijke ontwikkeling (o.a. organisatie, verantwoordelijkheid, communicatie) hebben bijgedragen tot dit uitzonderlijke gevoel van dankbaarheid. Peter, bedankt voor alle kansen die mij de afgelopen zes jaar aangereikt werden!

Wetenschappelijk werk wordt zelden door een enkel individu uitgevoerd. Een welgemeende dankbetuiging naar professor Wim Verbeke, die de voorzittersrol van de doctoraatsjury, bestaande uit professor Diederik Rousseau, professor Piet Verdonschot, professor Stijn Luca, doctor Peter van Puijenbroek en doctor Dries Landuyt, op zich nam. De gedetailleerde feedback heeft een onmiskenbare meerwaarde geleverd aan de structuur en de inhoud van dit doctoraat.

At the start, six years appear to be an eternity. However, when you have a great team by your side, time becomes relative and is perceived to fly by, while memories seem to cover a greater amount of time. A big thanks to our secretaries Sigrid and Marianne for taking care of both large and small technicalities and logistic issues. Sigrid, thanks for all the chats and shared moments of frustration ("*let's organise a conference*"), which made me wonder if you could be even more sarcastic (*eye roll*). A special thanks to my fellow assistants Niels, Emmanuel and Ilias for the fluent interactions on student-related matters. Niels, thanks for taking over during the final months of the writing process and the last-minute organisation of the online courses.

An extended thanks to the laboratory personnel Nancy, Gisèle, Emmy, Jolien and Marc for solving and supporting any practical or experimental issues. Nancy, thanks for the irreplaceable support before, during and after any sampling campaign. Of course, a huge thanks to all members of the Aquatic Ecology research group (present and past): Stijn, Long, Arne, Eurie, Daniel, Shewit, José, Lenin, Jawad, Tiptiwa, Tu Tri, Natalia, Sacha, Jana, Rubén, Naomi, Ratha, Selamawit, Tien, Gert, Pieter, … and the complete GhEnToxLab research group for the frequent end-of-the-week drinks and end-of-the-month activities.

Ook buiten de werkomgeving hebben meerdere mensen een (in)directe invloed gehad op de afronding van dit doctoraat. Een speciale bedanking naar Michaël, Kevin en Jochen voor de vele uren, dagen en verhalen die we delen en in de toekomst gaan beleven. Natalia, thank you for the trips and dinners we shared and for the promise of showing me Ecuador one day. Inge en Yannick, bedankt om meerdere malen te ageren als reisgezelschap. Een uitgebreide dankbetuiging aan Robson, Justine, Tom, Julie, Steffen, Tom, Lennart, Lies, Stien, Boris en vele anderen voor de fantastische filmavonden. Tevens een welgemeende bedankt aan Annelies (en bij uitbreiding alle vrijwilligers binnen Natuurpunt Gent) voor de gezellige babbels in het Natuurpuntcafé of in de Bourgoyen zelf. Het is vanzelfsprekend dat er nog veel mensen overblijven die invloed hebben gehad op het verloop van de afgelopen zes jaar. Het is een onmogelijke taak om al deze personen individueel te bedanken, maar aan alle personen die zich hierdoor aangesproken voelen: een dikke merci.

Finaal wil ik nog uitgebreid applaudisseren voor mijn ouders, die me doorheen de voorbije jaren steeds hebben ondersteund in al mijn beslissingen. Ik durf dan ook met zekerheid te stellen dat dit doctoraat er zonder hun steun niet geweest zou zijn. Deze bedanking geldt bij uitbreiding naar mijn volledige familie, die dit proces van iets verderaf hebben gevolgd en zich soms afvroegen wanneer ik echt zou beginnen werken in plaats van aan de universiteit te blijven studeren.

Aan iedereen/To everyone:

Bedankt!/Thank you!

"Take almost any path you please, and ten to one it carries you down in a dale and leaves you there by a pool in the stream."

Herman Melville in *Moby Dick* (1851)

Table of Contents

List of acronyms

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the receiver operating characteristic Curve
BBN	Bayesian Belief Network
BBNR	Biomass-based Nutrient Removal
BOD	Biochemical Oxygen Demand
CART	Classification and Regression Trees
CCI	Correctly Classified Instances
COD	Chemical Oxygen Demand
CRF	Conditional Random Forest
CTW	Constructed Treatment Wetland
CV	Cross-Validation
DT	Decision Tree
DW	Dry Weight
EC	European Commission
ES	Ecosystem Services
FL	Fuzzy logic
FN	False Negative
FP	False Positive
FWS	Free Water Surface
GAM	Generalised Additive Model
GAMM	Generalised Additive Mixed Model
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model

HSI	Habitat Suitability Index
HSM	Habitat Suitability Model
IAS	Invasive Alien Species
ICW	Integrated Constructed Wetland
kNN	k Nearest Neighbours
LMEM	Linear Mixed Effects Model
ls	Least Squares
MAR	Missing At Random
MCAR	Missing Completely At Random
MD	Missing Data
mF	missForest
MICE	Multiple Imputation via Chained Equation
MIR	Model improvement Ratio
MSE	Mean Squared Error
NMAR	Not Missing At Random
NRMSE	Normalised Root Mean Squared Error
PA	Presence-absence
PCA	Principal Component Analysis
PDP	Partial Dependence Plot
PO	Presence-only
RF	Random Forest
RGR	Relative Growth Rate
RMSE	Root Mean Squared Error
RNR	Relative Nutrient Removal
ROC	Receiver Operator Curve
RQ	Research Question
SDG	Sustainable Development Goal

SDM	Species Distribution Model
SMART	Specific – Measurable – Attainable – Relevant – Time-bound
SS	Suspended Solids
tN	Total Nitrogen
TN	True Negative
tP	Total Phosphorus
TP	True Positive
TSS	True Skill Statistic
UN	United Nations
VI	Variable Importance
WSP	Waste Stabilisation Ponds

List of symbols

$\boldsymbol{A}$	Matrix containing absence-related dummy scores
$\boldsymbol{C}$	Matrix containing Pearson correlation scores
$\boldsymbol{D}$	Data set
$\boldsymbol{D_{opt}}$	Data set with the highest number of complete data points
e	Numerical value of standard error for variable X
f_{MD}	Fraction of missing data
k_{cv}	Number of folds during cross-validation
k_{nn}	Number of neighbours as used in kNN
m	Mass
mtry	Number of variables to be considered in Random Forest nodes
n_rep	Number of repetitions during modelling
N_{inst}	Number of instances
$N_{inst,c}$	Number of instances in complete case analysis
$N_{inst,opt}$	Number of instances in $\boldsymbol{D_{opt}}$
nleaf	Minimum number of instances in a terminal node
nsplit	Minimum number of instances in a node to allow for splitting
ntree	Number of trees to be developed in Random Forest
N_{var}	Number of variables
$N_{var,c}$	Number of variables in complete case analysis
$N_{var,opt}$	Number of variables in $\boldsymbol{D_{opt}}$
N_z	Number of specified element z
$\boldsymbol{O}$	Matrix containing outlier-related dummy scores
P_z	z-th Percentile
$Q_{j,1}$	First quartile for variable j

$Q_{j,3}$	Third quartile for variable *j*
s	Numerical value of standard deviation for variable X
$\boldsymbol{s}$	Vector with standard deviations for all considered variables
Sn	Sensitivity, based on confusion matrix
Sp	Specificity, based on confusion matrix
x	Specific numerical value for variable X
$\boldsymbol{x}$	Vector with values for variable X
$\bar{x}$	Numerical average for variable X
$\overline{\boldsymbol{x}}$	Vector with average scores for all considered variables
α_a	Threshold for minimum number of exceedance
α_o	Threshold for acceptable number of outliers
β_o	Intercept in linear regression
β_j	Coefficient for term *j* in linear regression
Δ_{i-j}	Difference between metrics *i* and *j*
ε	Error in linear regression
κ	Kappa statistic, based on confusion matrix
τ_a	Threshold for absence selection
τ_c	Threshold for correlated variables
τ_i	Threshold for variable importance
τ_o	Threshold for outlier selection

Summary

Decreasing water availability and increasing global population cause tremendous pressures on the currently prevailing freshwater sources. It has become clear that synergies within freshwater management are required to simultaneously tackle the need for (i) improved water quality, (ii) increased water storage, (iii) efficient land use and (iv) preventing invasion impacts. Literature provides several options to tackle these threats in a simultaneous manner, including technological advancements and nature-based solutions. The latter entails the use of integrated constructed wetlands, which link the terrestrial with the aquatic system and house a variety of beneficial services to society. Yet, to avoid inefficient management, attention should be given to (i) identifying the biotic group with the highest potential to steer biotic development, (ii) determining locations suitable for species survival and (iii) defining the threat by invasive species.

Based on these issues, three themes were created to tackle these contemporary challenges and provide perspectives for decision makers and conservation managers. The first theme explores existing experiences in literature to create a conceptual framework. Secondly, attention is directed towards data-driven model development, with specific focus on the use and preparation of publicly available data followed by the development of abiotic habitat suitability models to infer species-specific habitat preferences. Lastly, the third theme applies autecological experiments to support both proactive and reactive management to mitigate invasion impacts. Finally, this work concludes with a comprehensive discussion and several promising perspectives.

Within the first theme, the literature review is divided into two main parts: (1) ecosystem services provided by wetlands and (2) advantages and disadvantages of data-driven habitat suitability models. First, **Chapter 2** describes how wetlands support sustainable development by providing pollutant reduction and by influencing biotic and abiotic interactions, ultimately concluding that macrophytes have a steering role regarding wetland community composition and functioning. Moreover, model selection, data quality assurance and controlled experiments are identified as attention points and addressed in subsequent chapters. For instance, **Chapter 3** describes the different steps within model development, including the conceptual framework, technique selection, model calibration and model validation. Advantages and disadvantages of five data-driven techniques (decision trees, generalised linear models, artificial neural networks, fuzzy logic and Bayesian belief networks) are discussed in a comparative context, along with various performance metrics to quantify model calibration and validation. The chapter concludes by recommending decision trees as a purely data-driven technique and thereby especially endorses the use of random forests.

The second theme discusses the development of macrophyte-specific habitat suitability models and starts by elaborating on data cleaning prior to model training. Within **Chapter 5**, the accuracy of four imputation techniques is discussed, being variable-specific mean, least square regression, *k* nearest neighbours and the *missForest* algorithm. A total of 720 data sets with artificially missing data is imputed and supports the overall conclusion that the *missForest* algorithm performs best. Subsequently, outliers, false absences and redundant variables are identified in **Chapter 6** by applying a range of potential threshold values. The results illustrate that model performance is clearly affected by data pre-processing and that a set of threshold values can be inferred to identify outliers (τ_o = 3), false absences (τ_a = 5 %), correlated variables (τ_c = 0.7) and irrelevant variables (τ_i = 10 %). The chapter concludes by indicating that serial data pre-processing improves model performance, while the presence of false absences in the test data deflates model validation scores. Lastly, building on these results, macrophyte-specific abiotic habitat suitability models are developed in **Chapter 7** thereby supporting relatively good discriminative and classification power. In addition, a set of major habitat descriptors is inferred along with their characteristic optimal conditions: temperature (> 17 °C), nitrate-N (0.5 mg·L^{-1} up to 1.5 mg·L^{-1}), oxygen (4 mg·L^{-1} up to 7 mg·L^{-1}), ammonium-N (0.3 mg·L^{-1} up to 0.5 mg·L^{-1}) and pH (7 up to 8.5). Yet, further fine-tuning of these ranges can be obtained via species-specific analyses.

Within the third and last theme, the focus is aimed towards avoiding the impact of invasive alien species by relying on proactive and reactive management. More specifically, **Chapter 8** introduces three indices to predict the invasive behaviour of the alien *Lemna minuta* in comparison with the native *L. minor*, being the functional response, the relative growth rate and the biomass-based nutrient removal. *L. minor* shows to remove more nutrients and develop more biomass, causing the chapter to conclude that the selected indices are insufficient to infer invasion potential. In contrast, reactive management is discussed in **Chapter 9** by exposing both *Lemna* spp. to nine different scenarios combining removal frequency ('none', 'low' and 'high') and biomass introduction frequency ('none', 'low' and 'high'). The results indicate slightly higher growth rates for *L. minuta* compared to *L. minor* and a negative feedback due to overcrowding. Moreover, it shows that total biomass benefits from species introduction and that dominance by the host species decreases in time. Both chapters highlight the need for more testing, considering their limited extrapolation power.

To conclude this work, **Chapter 10** summarises the findings of all chapters and illustrates the added value towards wetland conservation, with specific attention towards the pressures caused by climate change and invasive alien species. Moreover, alternative techniques for data collection, cleaning and analysis are introduced, along with the promising perspective of integrating field observations and experiments in order to merge the strengths of correlative and mechanistic modelling.

Samenvatting

De afnemende waterbeschikbaarheid en toenemende wereldbevolking zorgen voor een enorme druk op de nog beschikbare zoetwaterbronnen. Dit onderstreept het belang van synergiën in het zoetwaterbeheer om op een simultane wijze tegemoet te komen aan de vraag naar (i) verbeterde waterkwaliteit, (ii) toegenomen wateropslag, (iii) efficiënter landgebruik en (iv) verlaagde invasie-impact. De wetenschappelijke literatuur omvat verscheidene opties om deze uitdagingen aan te gaan, waaronder technologische vooruitgang en natuur-gebaseerde oplossingen. Laatstgenoemde omvat het gebruik van geïntegreerde artificiële wetlands, die gekenmerkt worden door het creëren van een link tussen het terrestrische en het aquatische systeem en het voorzien van een variëteit van gunstige diensten voor de maatschappij. Echter, om inefficiënt beheer tegen te gaan, dient er aandacht besteed te worden aan (i) de identificatie van de biotische groep met het hoogste potentieel om biotische ontwikkeling te sturen, (ii) het bepalen van de locaties die geschikt zijn voor het overleven van de beschouwde soorten en (iii) het definiëren van de bedreiging gecreëerd door invasieve soorten.

Gebaseerd op deze uitdagingen en aandachtspunten, werden drie thema's afgelijnd en behandeld om perspectieven te voorzien voor beleidsmakers en conservatoren. Het eerste thema omvat het beschrijven van de bestaande ervaring die in de literatuur vermeld worden teneinde een conceptueel kader te creëren. Vervolgens wordt er, gebaseerd op het ontwikkelde conceptuele kader, extra aandacht gegeven aan het ontwikkelen van datagedreven modellen. Deze ontwikkeling omvat een specifieke focus op het gebruik en de voorbereiding van publiek toegankelijke data, gevolgd door het ontwikkelen van abiotische habitatgeschiktheidsmodellen om geprefereerde habitatomstandigheden af te leiden. Het derde thema behandelt het gebruik van autecologische experimenten ter ondersteuning van proactief en reactief beheer met betrekking tot het mitigeren van invasie-impacts. Uiteindelijk sluit het werk af met een discussie en de identificatie van enkele veelbelovende toekomstperspectieven.

Binnen het eerste thema wordt het literatuuronderzoek opgesplitst in twee delen: (1) de ecosysteemdiensten die door wetlands worden voortgebracht en (2) de voor- en nadelen van datagedreven habitatgeschiktheidsmodellen. Allereerst wordt er in **Hoofdstuk 2** beschreven hoe wetlands bijdragen tot duurzame ontwikkeling door het voorzien van polluentverwijdering en door het beïnvloeden van verscheidene biotische en abiotische interacties. Er wordt besloten dat macrofyten een sturende rol hebben in de ontwikkeling van het beschouwde aquatische systeem en dat modelselectie, kwaliteitscontrole en gecontroleerde experimenten belangrijke aandachtspunten zijn. Deze elementen worden bijgevolg stapsgewijs in de volgende hoofdstukken behandeld.

Bijvoorbeeld, in **Hoofdstuk 3** worden de verschillende stappen in het modelleerproces beschreven, inclusief conceptueel kader, techniekselectie, modelkalibratie en modelvalidatie. De voor- en nadelen van vijf verschillende modelleertechnieken (beslissingsbomen, veralgemeende lineaire modellen, artificiële neurale netwerken, vage logica en Bayesiaanse netwerken) worden bediscussieerd in een vergelijkende setting, tezamen met meerdere performantie-indices om modelkalibratie en –validatie te beschrijven. Het hoofdstuk sluit af met het aanraden van beslissingsbomen als zuivere datagedreven modelleertechniek, met een specifieke vermelding van de *random forest* benadering.

Na de techniekselectie wordt data omtrent macrofytaanwezigheid verzameld, gekarakteriseerd en voorbereid voor het extraheren van patronen. Het opkuisen van data is relatief tijdsintensief en behandelt ontbrekende gegevens, extreme waarden, valse afwezigheden en redundante variabelen. In **Hoofdstuk 5** wordt dieper ingegegaan op de aanwezigheid van ontbrekende gegevens door de nauwkeurigheid van vier imputatietechnieken te beschreven, namelijk het variabele-specifieke gemiddelde, *least square* regressie, *k nearest neighbours* en het ensemble-gebaseerde *missForest* algoritme. De analyse omvat het artificieel verwijderen van data uit 720 datasets, gevolgd door imputatie en bepaling van de behaalde nauwkeurigheid. Er wordt besloten dat het *missForest* algoritme de hoogste nauwkeurigheid voorziet van de geselecteerde technieken.

Vervolgens worden extreme waarden, valse afwezigheden en redundante variabelen geïdentificeerd en geëlimineerd in **Hoofdstuk 6**, hetgeen resulteert in een analyse van de potentiële drempelwaarden. De resultaten illustreren dat modelperformantie beïnvloed wordt door het voorbehandelen van de beschikbare data en dat een set van drempelwaarden kan afgeleid worden om extreme waarden (τ_o = 3), valse afwezigheden (τ_a = 5 %), gecorreleerde variabelen (τ_c = 0.7) en irrelevante variabelen (τ_i = 10 %) te verwijderen. Het hoofdstuk sluit af met de observatie dat het voorbehandelen van data een positief effect heeft op modelperformantie, terwijl valse afwezigheden in de validatiedata kunnen leiden tot een lagere performantiescore.

Ter afsluiting van dit thema worden, op basis van deze resultaten, macrofyt-specifieke abiotische habitatgeschiktheidsmodellen ontwikkeld in **Hoofdstuk 7**, waarbij een goede discriminatie en classificatie bekomen wordt. Meer nog, een set van belangrijke habitatdescriptoren kan afgeleid worden, met karakteristieke optimale waarden voor macrofyt-aanwezigheid: temperatuur (> 17 °C), nitraat-stikstof (tussen 0.5 mg·L^{-1} en 1.5 mg·L^{-1}), zuurstof (tussen 4 mg·L^{-1} en 7 mg·L^{-1}), ammonium-stikstof (tussen 0.3 mg·L^{-1} en 0.5 mg·L^{-1}) en pH (tussen 7 en 8.5). Verdere detaillering van deze waardes kan bekomen worden via soort-specifieke analyses.

Binnen het derde en laatste thema wordt de focus gelegd op het vermijden en verminderen van de impact veroorzaakt door invasieve uitheemse soorten met behulp van proactief en reactief beheer. In **Hoofdstuk 8** worden drie indices voor het voorspellen van invasief gedrag voorgesteld en vervolgens toegepast om de uitheemse macrofyt *Lemna minuta* te vergelijken met de inheemse *L. minor*, namelijk de functionele respons, de relatieve groeisnelheid en een biomassa-gebaseerde nutriëntverwijdering. Binnen de bestudeerde nutriëntrange toont de inheemse *L. minor* een groter vermogen om nutriënten te verwijderen en biomassa te ontwikkelen, hetgeen veldobservaties tegenspreekt. Het hoofdstuk concludeert dat de gekozen indices onvoldoende zijn om het invasiepotentieel van invasieve uitheemse macrofyten te bepalen.

Vervolgens wordt in **Hoofdstuk 9** een reactief beheer toegepast en besproken, volgend op het blootstellen van beide *Lemna* spp. aan negen verschillende scenario's die verwijderingsfrequentie ('geen', 'laag' en 'hoog') en introductiefrequentie ('geen', 'laag' en 'hoog') combineren. De resultaten tonen een hogere groeisnelheid voor de uitheemse *L. minuta* vergeleken met de inheemse *L. minor* en een algemene afname in tijd door een toename in densiteit. Tevens wordt aangetoond dat de totale biomassa toeneemt door de introductie van biomassa en dat de biomassaverhouding tussen beide soorten afneemt in de tijd. De variatie in respons toont aan dat verdere studies aangeraden zijn om zowel proactief als reactief beheer te ondersteunen.

Om dit werk te eindigen, vat **Hoofdstuk 10** alle resultaten samen, waarmee de toegevoegde waarde naar het behoud en herstel van wetlands, met extra aandacht naar de druk die klimaatsverandering en invasieve soorten uitoefenen, wordt geïllustreerd. Alternatieve en supplementaire technieken voor dataverzameling, -reiniging en –analyse worden vermeld, tezamen met het potentieel van de verdere integratie van veldobservaties en experimenten om de sterktes van correlatieve en mechanistische modellen te combineren.

1

Introduction

Highlights

- Ongoing population growth, globalisation and climate change pressurise freshwater systems
- Wetlands provide various ecosystem services and are a valuable conservation option
- Application of models and autecological experiments is imperative to support freshwater management

Abstract

Water is essential to life on Earth, yet for centuries, running water has been considered a low-cost and energy-efficient disposal system for human settlements. With rising population levels, pressures on prevailing freshwater systems have increased rapidly, being additionally exacerbated by rising food and personal hygiene demands. As a response, the United Nations developed a calendar with 17 Sustainable Development Goals, to be completed by 2030, all of which are interconnected and allow for a certain degree of integration. It is clear that synergies within freshwater management are required to simultaneously tackle the need for (i) improved water quality, (ii) increased water storage, (iii) more efficient land use and (iv) inhibiting the effect of invasive alien species. To this end, wetland systems provide a potential starting point as they act as an important link between the terrestrial and aquatic system, while housing a variety of beneficial services to society. More specifically, attention should be given to (i) identify the biotic group with the highest potential to steer biotic development, (ii) determine locations suitable for species survival and (iii) define the threat by invasive alien species. These challenges are to be tackled by combining experience, experiments and models, with the latter increasingly relying on the growing field of artificial intelligence and publicly accessible data. By considering these issues, a series of research questions and objectives is defined and used for outlining the structure of this work.

1.1 Setting the scene

1.1.1 Global pressures and threats

Water is essential to life on Earth. It has been the basis for the first steps in the evolution process and has driven the development of human settlements for centuries. The uniqueness of such a resource being available throughout the world and originating from a huge, interconnected reservoir has not only supported the development of and revolutions within human history, but has also caused its abuse. For centuries, running water has been considered a low-cost and energy-efficient disposal system for human settlements, discharging liquid and solid wastes and relying on natural dilution and attenuation.

With relatively low historical population densities and wastewater mostly consisting of easily-degradable organic compounds, impacts on water quantity and quality due to extraction and discharge remained highly localised. This all changed with the start of the first Industrial Revolution during which water was viewed as a valuable energy source (Tvedt, 2010), seeding machine development, production proliferation and increased discharge of unwanted by-products. Continued reliance on the inherent attenuation power of nature caused uninterrupted and unregulated discharges, reflecting the 'Tragedy of the Commons': When something is freely accessible to all, it will be abused and overexploited until it becomes monetised and available to only a few (Hardin, 1968).

One of the main drivers underlying this tragedy is the uninterrupted growth of the global population, which crossed the virtual threshold of 1 billion around 1800 and has increased ever since. In 2020, the world population reached 7.8 billion people (Figure 1.1), while projections estimate the existence of 10 billion people by 2057 (United Nations, 2020). Rising food and personal hygiene demands pressurise the prevailing freshwater systems at a qualitative and quantitative level. More importantly, these pressures are being exacerbated by land use alterations and global climate change, causing disruptions of hydrological cycles on local, regional and global scales (IPCC, 2014; Verdonschot *et al.*, 2013). For instance, the first signs of these detrimental interactions are hard to ignore and include reports on the shrinkage of reservoirs and glaciers (Boomer *et al.*, 2000; Roe *et al.*, 2017), changes in frequency and intensity of rain patterns (Berg *et al.*, 2013) and increased faecal contamination of drinking water (UNESCO, 2017). These effects are expected to escalate in the future, thereby becoming either a cause for conflict or an opportunity for cooperation (Barnaby, 2009; Pearse-Smith, 2012; Shultz, 2003).

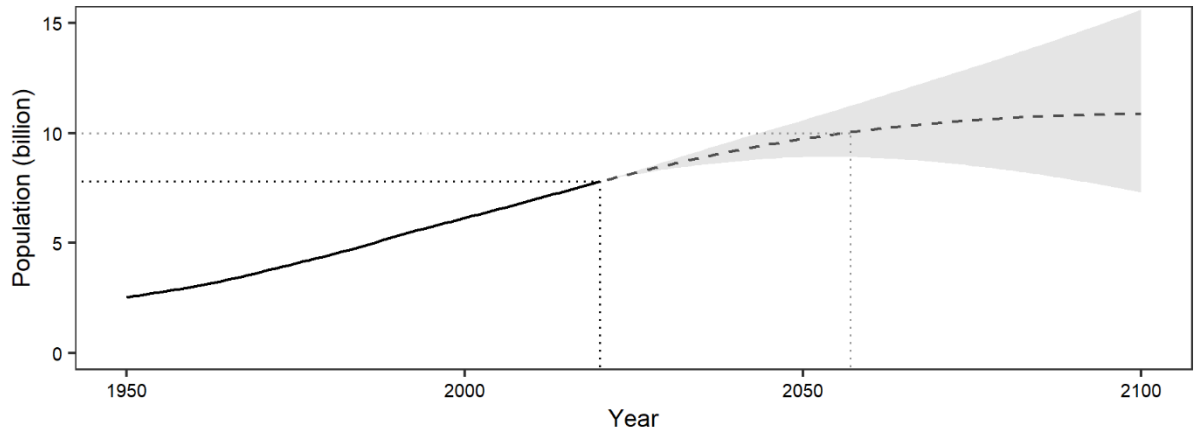

Figure 1.1: Evolution of the world's population. *In 1950, only 2.5 billion people lived on Earth, which has increased to 7.8 billion in 2020 (black dotted line) and is estimated to reach 10 billion by 2057 (grey dotted line). The inherent uncertainty of forecasts is depicted by a light grey ribbon surrounding the mean estimation (dashed black line) and reflects the difference caused by the considered fertility range. Data retrieved from United Nations (2020).*

Aside from the direct effects of altered hydrological cycles on the provisioning of water to society, also indirect effects are expected to increase due to modified water budgets within ecosystems (IPCC, 2014; UNESCO and UN-Water, 2020). Similar to human-oriented communities, natural systems depend on clean and abundant water for their development and to sustain their intrinsic complex interactions. Meanwhile, these systems provide a plethora of water-related ecosystem services (ES) that are intrinsically linked with water quality (e.g. purification) and water quantity (e.g. storage). These examples only represent a fraction of all the benefits that society can enjoy from natural systems, though their sustainable exploitation is challenged by delayed socio-economic acceptance and political impetus (Friberg *et al.*, 2017; Jähnig *et al.*, 2011). To improve understanding and awareness, additional distinction is made between provisioning (e.g. food, fibres), cultural (e.g. aesthetics, recreational) and supporting (e.g. nutrient cycling, soil formation) services (Millenium Ecosystem Assessment, 2005), along with several efforts to valuate ES (Costanza *et al.*, 2014). Still, each of these services relies on natural processes within a stable and functional ecosystem.

Unfortunately, pressures arising at the abiotic level threaten ecosystem structure and functioning, thereby negatively affecting the intensity of the provided ES. Habitat fragmentation, alterations in land use, chemical pollution and invasive alien species represent only a few underlying causes of the current biodiversity crisis (Harrison *et al.*, 2018; He *et al.*, 2019), with species extinction rates and reductions in wildlife populations reaching unprecedented levels (Vitousek *et al.*, 1997; WWF, 2018). These observations call for immediate management measures, despite the inherent complexity of stressor interactions.

1.1.2 Responding to freshwater pressures

Clean and abundant water acts as a cornerstone for socio-economic development, making it an undeniable human right. With the ongoing pressures in mind, the United Nations created the Sustainable Development calendar for 2030 as a sequel to the Millennium Development Goals, which ended in 2015. Within the new framework, 17 Sustainable Development Goals (SDGs) have been identified to support future development with attention to social, economic and environmental aspects (United Nations, 2015).

All SDGs are closely linked to each other, yet focus on different aspects of sustainable development. Issues related to freshwater are the main topic in SDG 6 (Clean water and sanitation) and SDG 15 (Life on land) and, despite being part of different SDGs, allow to be partially tackled simultaneously. For instance, treatment of wastewater prior to its discharge reduces the pollutant load into the environment, allowing for lower purification costs (into clean water) and providing less pressure on the biotic communities within the water. With 2.1 billion people lacking access to safe drinking water and about 80 % of all wastewater entering the environment without treatment, the urgency to accelerate efforts within water-related goals is unambiguously clear and consolidates the declaration of the International Decade for Action on Water, running from 2018 until 2028 (UNESCO, 2017; United Nations, 2020; WHO/UNICEF, 2017).

The abovementioned local anthropogenic activities negatively affect the provisioning of sufficient and qualitative water, while increased travel and trade at a global scale continuously transport organisms outside their native range, both intentionally and accidentally (Perrings *et al.*, 2002). The introduction of an alien organism (see Box 1.1 for related terminology) into a suitable environment can both improve and threaten community composition, while additionally affecting economic activities and human health (Born *et al.*, 2005; Pejchar and Mooney, 2009; Strayer, 2010). For instance, the megafauna in Australia are all introduced species, including species that are at risk or extinct in their native range (Lundgren *et al.*, 2018). In contrast, the invasion of the zebra mussel (*Dreissena polymorpha*) has caused a widespread occurrence within the Mississippi basin and has led to the clogging of multiple water intake pipes (Ludyanskiy *et al.*, 1993). Similarly, the floating water hyacinth (*Eichhornia crassipes*) has become a common inhabitant of tropical lake systems around the world, yet blocks several aquatic transport routes (Villamagna and Murphy, 2010). The resulting ecological and economic impacts have been so severe that both species earned a spot in the IUCN List of the 100 worst invasive species (IUCN, 2019). Currently, border control represents the most-developed proactive management measure to impede the introduction of alien species, though the increasing number of reports on alien species indicates a need for alternative and supplementary measures in order to avoid expensive (and often ineffective) eradication programs (Early *et al.*, 2016; Williams and Grosholz, 2008).

Box 1.1: Terminology biological invasions

A variety of terminology and definitions is used within the field of biological invasions, which impedes transparent communication and challenges decision-making (Blackburn *et al.*, 2011; Colautti and MacIsaac, 2004). For instance, Colautti and MacIsaac (2004) consider a species to be invasive when depicting a significant spread in the new geographical range, while Davis and Thompson (2000) state that a severe impact is required prior to being considered invasive. While redefining the complete field of biological invasions is beyond the scope of this work, it is clear that overall transparency can be improved by defining the terminology to be used throughout the following chapters.

Each species is characterised by a specific geographical range in which it naturally occurs, survives and reproduces. Species occurring within their natural geographic range are referred to as **native species**. In contrast, species can be transported outside their native range by anthropogenic activities and be introduced in a new environment. These species are referred to as **alien species**, **exotic species**, **non-native species** or **non-indigenous species**. When species have the tendency to completely colonise and outcompete the prevailing populations after their arrival in a new site (be it within or outside the native range), they are considered to be **invasive**. Hence, within a specific geographical area, both native and alien species can display invasive behaviour.

For a species to be classified as an Invasive Alien Species (IAS), it has to go through several stages and overcome multiple barriers, which has been summarised in various conceptual frameworks. For instance, Blackburn *et al.* (2011) combine the individual-based approach of Richardson *et al.* (2000) and the population-based approach of Williamson and Fitter (1996) in a unified framework with the following four stages: (1) **Transport**, (2) **Introduction**, (3) **Establishment** and (4) **Spread/Colonise**.

Each of these stages is characterised by one (or more) barrier(s). More specifically, geographical restrictions limit the number of species that will be transported, while cultivation/captivity impedes the introduction of a selected set of species into a new environment. The latter represents an optional barrier, as many other species (e.g. plants, fungi, invertebrates) have the capacity to be unintentionally transported and directly introduced in the new environment. Following introduction, species need to be able to survive and reproduce within their new environment in order to become an established and self-sustaining population. Lastly, successful spread is reached when overcoming the dispersal and subsequent environmental barrier. With each barrier, a fraction of alien species is lost and considered unfit to significantly impact native communities in the long term (Blackburn *et al.*, 2011; Williamson and Fitter, 1996).

The identification of these pressures and underlying driving forces helps in pinpointing bottlenecks and developing responses to mitigate impacts, following the conceptual DPSIR (Driving Force – Pressure – State – Impact – Response) approach (Vannevel, 2018; Verdonschot *et al.*, 2013). Increased awareness on the state and the intrinsic value of natural ecosystems has kick-started research on the applicability and potential of ecosystem management (see Box 1.2 for related terminology) to counter anthropogenic pressures. So far, positive results have been obtained for projects aiming to improve abiotic conditions by implementing re-meandering, breaking down weirs and installing wastewater treatment plants (Jähnig *et al.*, 2011; Lorenz *et al.*, 2009).

In contrast, pilot studies on biological restoration following these abiotic improvements have provided mixed results due to the high spatiotemporal and biological complexity of natural ecosystems (Hilderbrand *et al.*, 2005; Verdonschot *et al.*, 2013). For instance, dispersal limitations, biotic resistance and the absence of a proper ecosystem engineer are only a few processes that can cause a significant temporal delay to reach the project-specific goals. To counter these delays, manual introduction can help accelerating natural succession, though relies on species-specific assessment of habitat suitability and significant monetary investments (Lu *et al.*, 2012; Zhang *et al.*, 2017).

From this, it is clear that cooperation and integration of individual freshwater management activities is required to simultaneously tackle the need for (i) improved water quality, (ii) increased water storage, (iii) more efficient land use and (iv) limiting the impact of invasive alien species. Such synergies occur naturally near the border of existing ecosystems, ecotopes and habitats by locally integrating and fine-tuning characteristics of all contributing components (Banks-Leite and Ewers, 2009). The resulting complexity and extensiveness tend to vary greatly in function of the severity in change, ranging from highly abrupt (e.g. rocky cliffs, glacial lakes) to smooth (e.g. local topographic depressions, estuaries). Moreover, due to the recurring difficulty in defining a clear border between neighbouring habitats, the transition zone can be relatively wide and cover an additional habitat type (Banks-Leite and Ewers, 2009; Strayer *et al.*, 2003).

For instance, wetlands are characterised by a smooth transition between the aquatic and terrestrial system and tend to develop differently depending on the prevailing environmental conditions. Consequently, a variety of subtypes exists, including fens, bogs, peatlands, marshes, mangroves and swamps, which share only the presence of a hydric soil as a common factor and leave proper delineation open for discussion (Dodds and Whiles, 2010; Gopal, 2016; Keddy, 2010; Kivaisi, 2001). In fact, wetlands combine aquatic and terrestrial characteristics and thereby provide a potential starting point to look for synergies and convey ecological conservation (Junk *et al.*, 2014; Kingsford *et al.*, 2016). Throughout the remainder of this work, wetlands will be considered as 'systems with a continuously waterlogged soil'.

Box 1.2: Terminology ecosystem management

Managing ecosystems entails all anthropogenic actions that directly and indirectly affect ecosystem composition and functioning, ultimately aiming to reach human-defined goals. Due to this variety of available actions, it is considered useful to introduce management-specific terminology and what it entails with respect to goal definition and field activities.

A first distinction can be made between **preservation** and **conservation**. Preservation aims at maintaining ecosystems in their pristine state, without society experiencing economic benefits. Conservation is less strict and aims at improving natural conditions (e.g. landscapes, biodiversity) while simultaneously considering potential benefits (i.e. ecosystem services) to and cooperation with society (Sarkar, 1999). Hence, conservation can be considered as more complex than preservation as it requires more fine-tuning with a human element.

Conservation plays at a large spatial scale and underlies several international agreements, including the definition of RAMSAR sites, Aichi targets and sustainable development goals (CBD, 2020; United Nations, 2015). To reach conservation at a large scale, small-scale activities and implementations are of importance. These can be broadly classified into (1) **Protection**, (2) **Restoration** and (3) **Construction**. Protection aims at the maintenance of an ecosystem and preventing its decline by eliminating pressures, without causing an increase in area of the considered system (sometimes referred to as **mitigation**) (Jackson *et al.*, 1995).

Restoration focuses on the improvement of the prevailing conditions in order to support natural development towards natural or historical conditions (Jackson *et al.*, 1995; Jackson and Hobbs, 2009). Depending on the author, restoration efforts can be considered in a broader sense and additionally include actions that (i) improve specific ecosystem functions (**enhancement**) and (ii) re-create structure and/or functioning without aiming towards historical conditions, distinguishing between a relatively high (**reclamation**) and low (**rehabilitation**) similarity with the reference ecosystem (Aronson *et al.*, 1993; Harris *et al.*, 2006; Jackson *et al.*, 1995; Jackson and Hobbs, 2009).

Lastly, construction supports the premise that ecosystems can be built at locations where they never occurred before in order to mitigate losses elsewhere or to locally improve the production of ecosystem services. Often, these actions are also referred to as representing **creation**, **reallocation** or **establishment** of the preferred artificial system (Aronson *et al.*, 1993; Jackson *et al.*, 1995).

1.1.3 Wetlands as a starting point

Natural wetlands have been around for as long as humans stroll around the world, but their area has decreased ever since (Kingsford *et al.*, 2016). Population growth, increased urbanisation and industrial development are only a few of the driving forces that have steered this downward trend and that have, along with other land transformations, caused the loss of at least 50 % (and potentially up to 87 %) of all wetland area around the world (Davidson, 2014; van Asselen *et al.*, 2013; Vitousek *et al.*, 1997). With wetlands providing key environmental processes and being identified as one of the highest-valued habitats per unit area (Costanza *et al.*, 2014; Millenium Ecosystem Assessment, 2005), they inherently affect local and regional ecosystems at the abiotic and biotic level (Zedler, 2003). Hence, wetland protection and restoration are of key importance to avoid future degradation and to regain lost functions on land (SDG 15 – Life on land) (Kingsford *et al.*, 2016; United Nations, 2015).

Artificial wetlands help to mitigate these losses by mimicking natural wetland conditions (Kadlec and Wallace, 2008; Scholz *et al.*, 2007), though are frequently built as single-purpose systems, including food production (e.g. rice paddies), flood protection (e.g. controlled flood areas) and pollution mitigation (e.g. reed beds). Application of the latter to fight point and diffuse pollution sources has received increased attention throughout the past fifty years, focusing on design, applicability, resilience, type of substrate and vegetation (Auvinen *et al.*, 2016; Karathanasis *et al.*, 2003; Kivaisi, 2001; Park and Polprasert, 2008; Rousseau *et al.*, 2004a; Vymazal, 2010). Due to their low capital and maintenance costs, constructed treatment wetlands (CTWs) represent a viable pollution mitigation measure in remote areas (Kivaisi, 2001; Vymazal, 2011a; Zhi and Ji, 2012), providing sanitation and cleaner water (SDG 6 – Clean water and sanitation) (United Nations, 2015).

Throughout the last two decades, multi-purpose designs that combine biodiversity increase and pollutant removal have been simultaneously introduced as the 'Water Harmonica' concept (Kampf and Claassen, 2004) and the 'Integrated Constructed Wetland' (ICW) concept (Scholz *et al.*, 2007). Within these concepts, CTWs are designed to provide both pollutant removal and landscape integration, while establishing a range of habitats to support increased biological diversity (Boets *et al.*, 2011; Harrington and McInnes, 2009; Scholz *et al.*, 2007). Nevertheless, reports on the combined pollutant reduction and biodiversity boost provided by ICW systems remain limited as most studies focus on the water treatment function (Becerra-Jurado *et al.*, 2012; Benyamine *et al.*, 2004; Hansson *et al.*, 2005). Further studies are therefore crucial to narrow the resulting gap between the conceptual framework and practical implementation.

1.2 Delineation of the study and research objectives

1.2.1 Identification of the working field

Integrated constructed wetlands (ICWs) rely on a complex interplay of physical, chemical and biological processes that deserve attention during decision-making and prior to implementing on-site measures. More specifically, the success of ICWs (either after construction or restoration) ultimately depends heavily on (1) the integration in its surrounding, (2) the degree of pollutant removal and (3) the resulting augmented biodiversity. With the current freshwater biodiversity crisis in mind (Harrison *et al.*, 2018; He *et al.*, 2019), the majority of this work is dedicated to the biodiversity potential of ICWs, without completely ignoring the physical and chemical aspects.

The biological response to the prevailing abiotic conditions and dynamics can be inferred from experiments, models or a combination of both. More importantly, both data sources entail a continuum that ranges from a simplified to a highly complex approach. For instance, experiments can be performed under controlled laboratory conditions with a single treatment factor, though can be as complex as restoring hydraulic conditions and assessing the difference in species richness over time. Similarly, models to infer species-specific habitat suitability and distribution patterns can be purely data-driven (empirical) or completely mechanistic (process-based), yet the design and application of all models is greatly determined by their intended usage.

This variety in experiment and model complexity requires a further delineation of the working field considered throughout this work. Given the increasing importance of environmental data science in decision-making and the growing amount of publicly-available occurrence data sets (Gibert *et al.*, 2018a; Maldonado *et al.*, 2015), it was decided to work with data-driven modelling techniques. These models allow for inferring species-specific habitat preferences, though tend to be challenged by a lack of data or by limited integration of species dynamics. This is especially the case for rare and alien species, which advocates the use of simplified experiments to infer and forecast species-specific behaviour. In short, both models and experiments are considered and applied to support the biotic restoration and construction of ICWs.

Aside from this conceptual delineation, several boundary conditions require specification prior to identifying the knowledge gaps and associated study objectives. Firstly, the physical design is assumed to promote relatively high hydraulic retention times and to represent inclined banks that allow for a gradient in water depth (and associated microhabitats). Secondly, the chemical conditions mainly represent a wastewater polishing stage and are, therefore, assumed to reflect elevated nutrient levels. Thirdly, the geographic location is focused on Belgium and the Netherlands and assumes a similar climate (i.e. no important steering climatic variables).

1.2.2 Research objectives

The conceptual delineation of the study area (see Section 1.2.1) creates a transparent foundation for outlining the practical research objectives of this work. These objectives help to link and streamline individual studies and can be easily divided into three major themes: (1) literature review, (2) data-driven modelling and (3) autecological experiments.

To start, literature provides an essential basis to narrow the practical working field further. More specifically, ICWs have already been introduced in Section 1.1.3, though deserve a more in-depth description of the various chemical processes and biotic interactions that take place within. Similarly, a variety of data-driven modelling techniques exists, which merits a detailed qualitative comparison prior to technique selection. Specific research objectives related to the literature review on ICWs and modelling techniques are provided in Section 1.2.2.1.

Secondly, data-driven modelling is not limited to technique selection, but also includes data cleaning and pre-processing in order to improve the quality of the data. This is especially the case when dealing with publicly available data, as these often contain noise and impure information (Maldonado *et al.*, 2015). The geographical delineation of the study allows the use of the Limnodata Neerlandica (Knoben and van der Wal, 2015), which is characterised by a relatively high spatiotemporal coverage. The structure of this data set supports the development of models trained with presence-absence data, which narrows the number of techniques to be considered in the first theme. Specific research objectives related to data-driven modelling are introduced in Section 1.2.2.2.

Thirdly, experiments provide valuable information when insufficient data is available for model development. The conceptual framework entails open water systems with elevated nutrient levels and are, similar to other freshwater systems, exposed to the introduction of alien species. Only a fraction of the introduced alien species survives (see Box 1.1), though these survivors can drastically affect ecosystem structure and functioning. Therefore, alien species with a negative effect on native species are best known in advance to support proactive management. In contrast, when such a species is already present, reactive management is needed to reduce its impact. Data-driven models are generally incapable of providing appropriate answers to these questions, which highlights the importance of experiments. Specific research objectives related to these autecological experiments are summarised in Section 1.2.2.3.

Throughout this work, these three themes are dealt with in the presented order, and subdivided in a series of research questions and related objective. For each theme, a visual representation is provided, along with the identification of the chapter dealing with the objective(s).

1.2.2.1 Theme 1: Exploring experiences

When working with integrated constructed wetlands, identification of a biotic group that represents habitat modifiers is recommended as they shape and transform the local ecosystem. Hence, the first research question is summarised as: "*Which biotic groups are relatively strong habitat modifiers?*" (Figure 1.2). An answer to this question is obtained by creating an overview of how biotic groups interact in shallow eutrophic freshwater systems (Objective 1.1) and determining which group has a relatively large impact on both the abiotic conditions and biotic community (Objective 1.2).

Secondly, a detailed description of the system under study is essential to construct the overall framework. Therefore, the second research question within this theme entails: "*What hampers implementing Integrated Constructed Wetlands (ICWs)?*" (Figure 1.2). An answer to this question is obtained by summarising wastewater treatment performance of constructed wetlands (Objective 1.3) and elaborating on the desired functions to identify current knowledge gaps (Objective 1.4).

Lastly, species occurrence is highly dependent on the prevailing abiotic conditions and biotic interactions, which can be combined in a modelling framework. Yet, as the number of available techniques increases rapidly, the following research question remains: "*What options exist for correlative habitat suitability modelling?*" (Figure 1.2). By comparing a selection of modelling techniques (Objective 1.5) and describing the overall modelling framework (Objective 1.6), an answer to this question is provided.

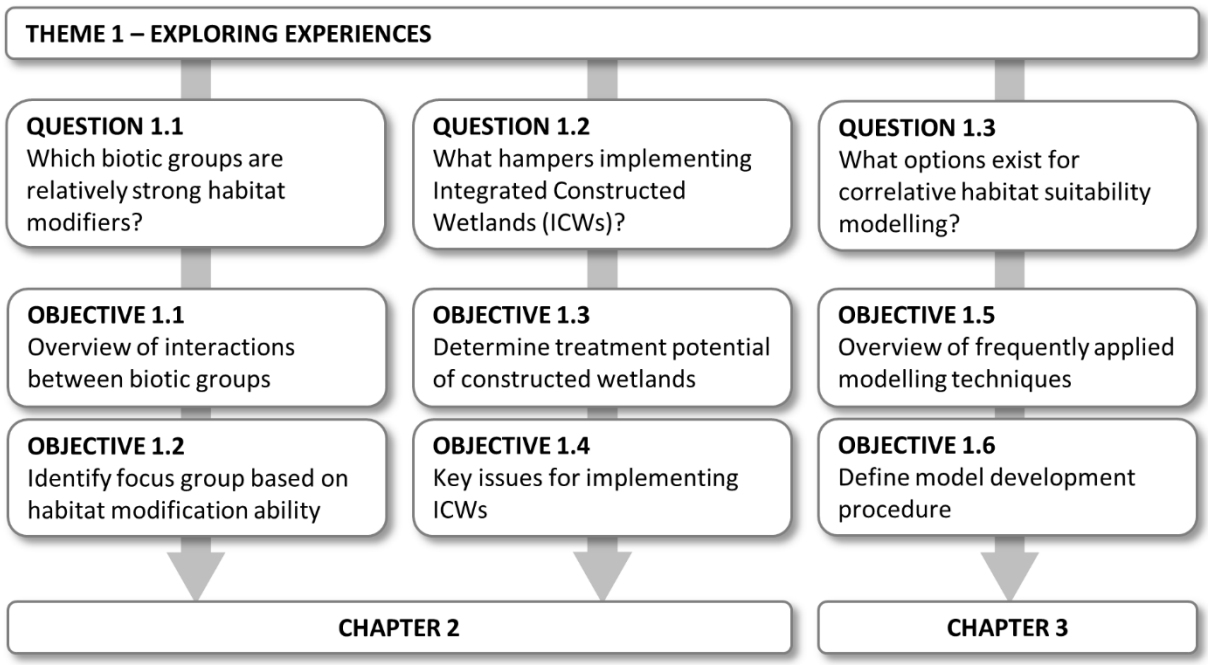

Figure 1.2: Content of the first theme, including research questions and underlying objectives. *Research question 1.1 and 1.2 are discussed in Chapter 2, while research question 1.3 is discussed in Chapter 3 (see further).*

1.2.2.2 Theme 2: Model development

When developing ecological models, natural processes and interactions are simplified to ease interpretation by complexity reduction (Wilson *et al.*, 2011). Therefore, model results should be interpreted with care, especially when publicly available data is used. This real-world data is generally in need of cleaning to improve the overall information density prior to being used, thereby positively affecting model fit and related results (Maldonado *et al.*, 2015; Zhang *et al.*, 2003). Hence, the first research question of this theme can be summarised as: "*How to prepare the available data to improve model performance?*" (Figure 1.3). An answer to this question is obtained by identifying and applying a technique to deal with missing data (Objective 2.1), along with exploring data cleaning procedures and related threshold selection to increase the information content (Objective 2.2). With data and time being valuable aspects during modelling, related gains or losses will be juxtaposed with changes in accuracy.

Subsequently, the pre-processed data act as information source for the development of predictive models in order to identify those locations that will benefit from artificial introduction. Moreover, it also allows to identify locations that remain unsuitable for native species, yet suitable for invasive alien species. Therefore, the second research question of this theme entails: "*How applicable is the selected modelling technique?*" (Figure 1.3). By developing species-specific models (Objective 2.3), derive species-specific habitat descriptors (Objective 2.4) and applying these models within a management framework (Objective 2.5), an answer to this research question is obtained.

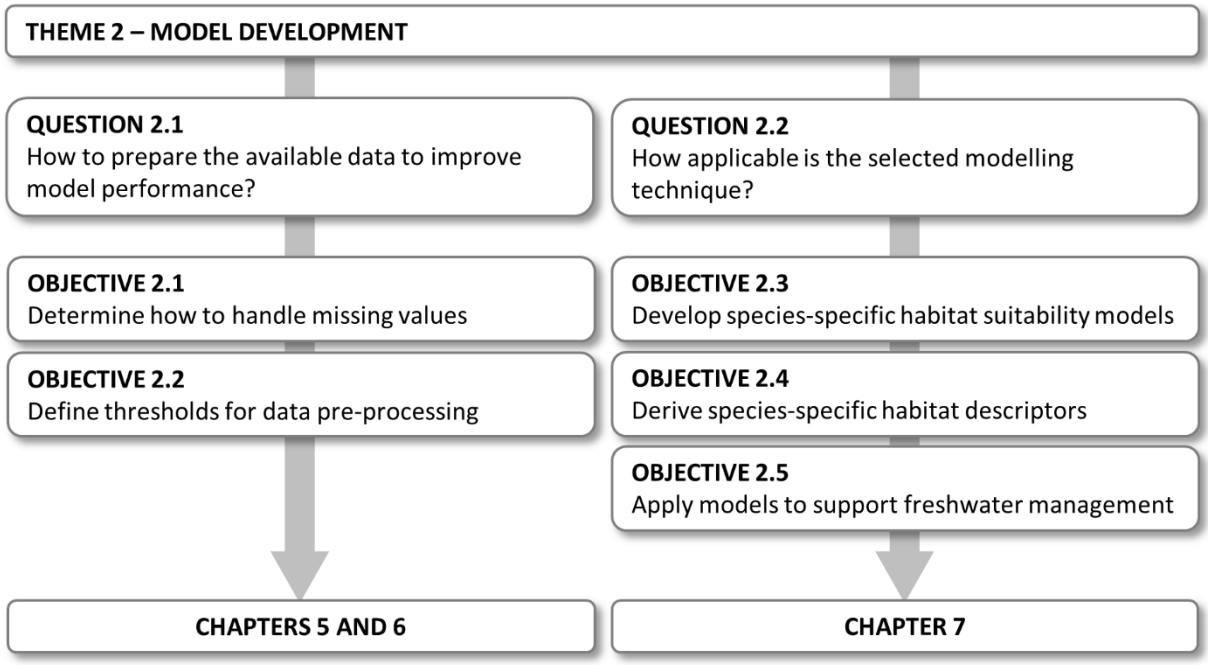

Figure 1.3: Content of the second theme, including research questions and underlying objectives. *Research question 2.1 is discussed in Chapters 5 and 6, while research question 2.2 is discussed in Chapter 7 (see further).*

1.2.2.3 Theme 3: Autecological experiments

Management of invasive alien species is significantly supported by the development of correlative habitat suitability and species distribution models (Boets *et al.*, 2010; Gallardo *et al.*, 2012). Yet, the data-driven nature of most modelling techniques relies on the presence of these non-indigenous species within the non-native range, thereby hampering their application to support proper proactive management. Hence, the first research question within this theme is summarised as: "*Can functional traits be used to infer invasive behaviour of alien species?*" (Figure 1.4). An answer to this question is obtained by selecting traits according to the SMART guidelines (Specific – Measurable – Attainable – Relevant – Time-bound) (Objective 3.1), followed by the comparison of field observations with the achieved trait results (Objective 3.2).

Secondly, management of freshwater sites that have been invaded by an alien species can be based on developed habitat suitability or species distribution models. Yet, only a fraction of modelling techniques is able to substantially include temporal dynamics, which illustrates a major drawback of model-based management. Moreover, it represents the basis of the second research question within this experiment-based theme: "*How does partial biomass removal affect species productivity?*" (Figure 1.4). By experimentally determining biomass production and ratio under different pressures (Objective 3.3) and comparing the response of a native and alien population (Objective 3.4), an answer to this research question is obtained.

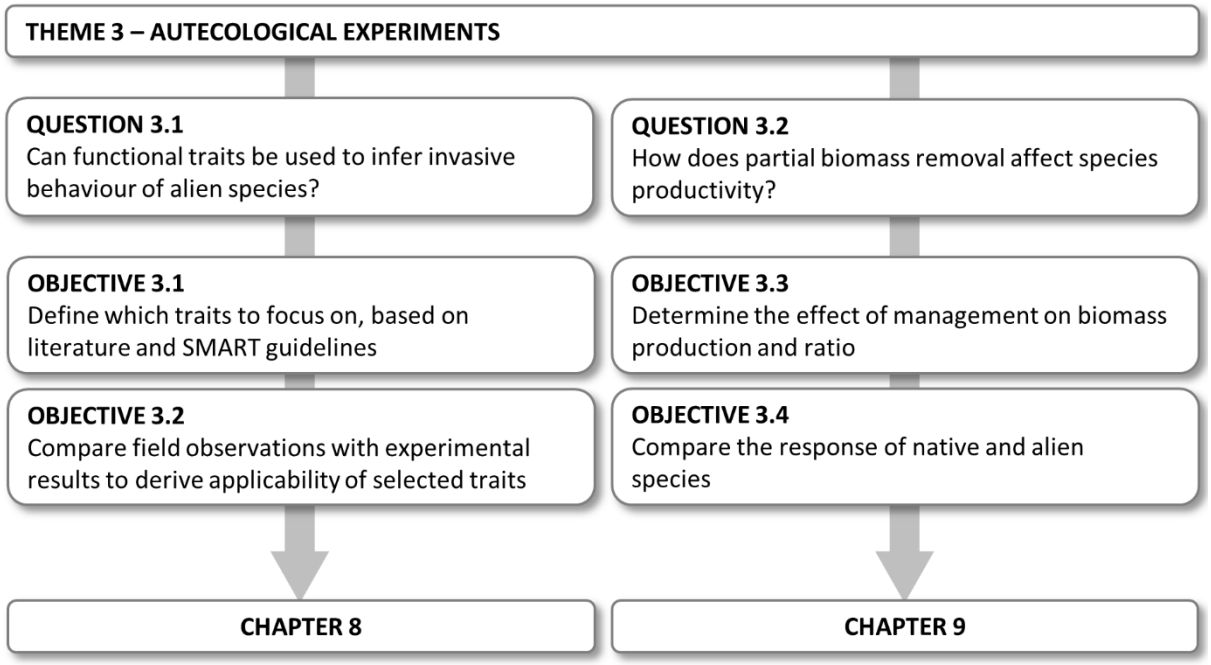

Figure 1.4: Content of the third theme, including research questions and underlying objectives. *Research question 3.1 is discussed in Chapter 8, while research question 3.2 is discussed in Chapter 9 (see further).*

1.3 Thesis roadmap

To answer the abovementioned research questions, several steps are taken, analysed and discussed throughout this thesis, separated in 10 different chapters. The content of most chapters and how they tackle a specific research question has already been shortly presented in Figure 1.2, Figure 1.3 and Figure 1.4, though merits a more elaborate description to reflect the overall structure and continuity of this work. Throughout the following paragraphs, the content of each chapter is introduced along with its relevance and contribution to the aforementioned research questions, which is ultimately summarised in a comprehensive scheme (see Figure 1.5).

Within this first chapter, the general background of the study is provided to introduce the societal relevance and scientific necessity of this work. Attention is given to the importance of water in society, as well as the value of aquatic ecosystems and the provided services. Due to the connection between aquatic and terrestrial systems, wetlands are considered to be a prominent starting point for combining water treatment, storage and purification along with a positive note towards the improvement of terrestrial biodiversity and river conservation. Based on this starting point, a series of research questions with underlying objectives are identified to steer and frame all subsequent chapters.

In **Chapter 2**, additional focus is given to the concept of Integrated Constructed Wetlands (ICWs) and the identification of interactions among several biotic groups within shallow eutrophic freshwater systems. Specific attention is given to pollutant removal within constructed wetlands and the structuring role of macrophytes in many aquatic systems. The chapter concludes with a summary of key issues that need further investigation to support the implementation of ICWs as a multi-purpose technique dealing with water pollution and biodiversity improvement. The majority of subsequent chapters elaborates on one (or more) of the key issues identified here.

Next, **Chapter 3** dives into the world of data-driven habitat suitability and species distribution models by discussing the advantages and drawbacks of five modelling techniques. Application of these techniques within ecosystem management is illustrated by means of examples, while highlighting the different steps and approaches to be considered during model development. The chapter includes an introduction to decision trees, generalised linear models, artificial neural networks, fuzzy logic and Bayesian belief networks as well as a description on model conceptualisation, requirements, calibration and evaluation. Moreover, it provides an introduction to the potential of and criticism on ecological modelling to support environmental management. The chapter concludes with the endorsement of a promising data-driven modelling technique.

Subsequently, **Chapter 4** provides a general description of the data and modelling technique underlying the following chapters. An in-depth description of the data set is provided (spatiotemporal coverage, number of instances, number of explanatory variables, number of species, missing data, …), as well as a more detailed discussion of the selected model algorithm. Moreover, the methodology and experimental design behind the applied data cleaning are touched upon. Finally, the focus species of the autecological experiments are introduced.

In **Chapter 5**, the first step in cleaning the available data is performed. More specifically, different approaches to deal with missing data are introduced to avoid the traditional information loss caused by removing the incomplete instances from the data set. Four techniques replacing the missing value by a data-derived value (i.e. 'data imputation') are discussed in more detail, being *mean value*, *least squares*, *k-nearest neighbours* and *missForest*. The application of each technique to a range of differently-sized data sets provides a conclusion on the most accurate technique, while mentioning computation time as a side aspect during method evaluation.

Following data selection and preparation, **Chapter 6** discusses the potential of further data pre-processing in concert with the selected algorithm, while identifying a lack of clear guidelines. The removal of redundant variables and potentially faulty instances is expected to reduce overall data complexity, to improve model performance and to decrease computation time. Throughout the chapter, four techniques are dealt with in more detail: (i) instance removal based on outliers, (ii) instance removal based on false absences, (iii) variable removal based on correlation score and (iv) variable removal based on variable importance. The chapter concludes with a statement on the effects of data pre-processing on model performance and provides a suggestion for further research.

Next, **Chapter 7** builds further on the findings of Chapter 5 (imputation technique) and Chapter 6 (data pre-processing) and combines them in the development of species-specific abiotic habitat suitability models. For each species, model fit is improved by optimising hyperparameter settings and comparing final model performance with baseline and null model performance. Based on these models, species-specific variable importance is derived, while additionally providing the opportunity to assess the effect of different management scenarios. Finally, the chapter concludes with an overview of the identified steering variables and how management effects depend on the starting conditions within the system under consideration, while highlighting that controlled experiments can improve understanding of the dynamic interactions occurring within ecosystems.

After the model-based approach, **Chapter 8** considers the application of controlled laboratory experiments to define invasive behaviour of an alien species. Two highly similar species are exposed to a range of nutrient concentrations under optimal growing conditions, while a selection of functional traits is being monitored. Based on ecological theory, it is expected that an invasive species exhibits higher performance either at the level of resource intake or biomass production. Finally, a conclusion on the applicability of the selected traits for inferring invasive potential of alien species is provided.

In contrast, **Chapter 9** elaborates on the post-establishment phase, where an invasive alien species is continuously introduced in a new environment and threatens native populations. Management of these native species by means of harvesting can disturb natural conditions and benefit the alien species. Meanwhile, biomass removal of an invasive alien species can help creating opportunities for the re-colonisation by (a) native species. Within this chapter, a dynamic interaction of management and introduction pressure is applied on the same species from Chapter 8 and is studied to infer the potential detrimental effects of management without prior study. Based on these observations, management suggestions are formulated to conclude this chapter.

Finally, **Chapter 10** combines all the observations into a general discussion, followed by a conclusion for future freshwater management. Within this chapter, the answers to the research questions and objectives identified in the first chapter are summarised and re-framed in a bigger story. More specifically, the chapter discusses the application of the suggested modelling technique in combination with the appropriate data pre-processing and hyperparameter optimisation and subsequently couples back with the need for restoring and constructing wetlands. Moreover, attention is given to the threat posed by invasive alien species and how performed experiments help in identifying solutions and challenges for both proactive and reactive management. Ultimately, the chapter concludes with an introduction to the future perspectives of ecosystem modelling and invasive species management.

From this, it is clear that all chapters are linked and provide a linear story throughout the whole thesis. Each chapter tackles a specific research question and the underlying objectives (see Figure 1.2, Figure 1.3 and Figure 1.4), while often building on previous chapters. A complete overview of this work and how the chapters are linked, is provided in Figure 1.5.

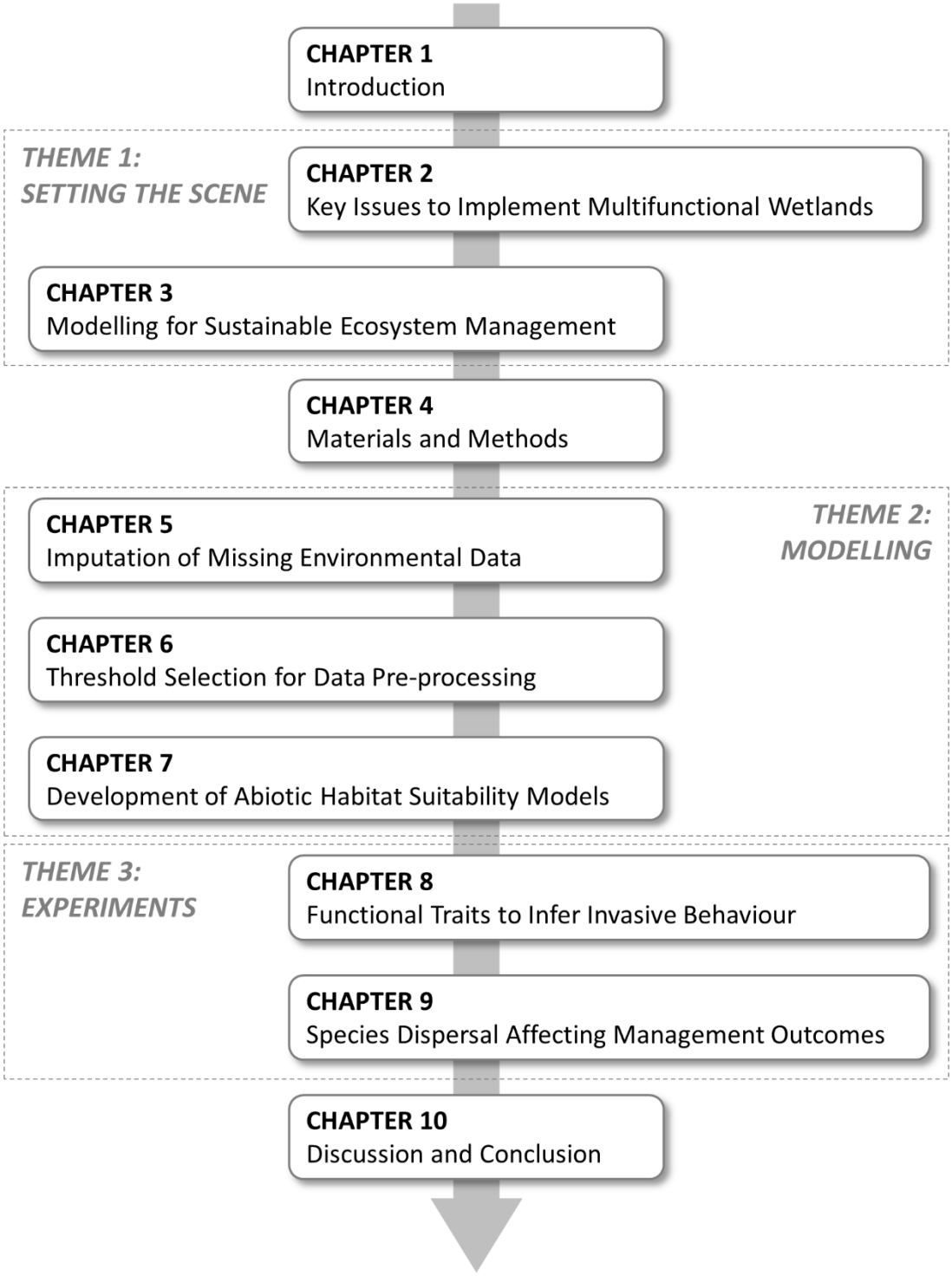

Figure 1.5: Roadmap of this thesis. *Three main themes can be identified: (1) Summarising available information by means of literature review (Setting the scene); (2) Optimise the available data via imputation and pre-processing and develop habitat suitability models (Modelling) and (3) Perform controlled experiments to complement the models (Experiments).*

2

Key issues for implementing artificial multifunctional wetlands[1]

Highlights

- Macrophytes support habitat structuring and positive biotic interactions
- Habitat suitability models provide potential to identify suitable species
- Chemical treatment and biotic improvement can be combined

[1] This chapter is redrafted from Van Echelpoel, W.; Donoso, N. and Goethals, P. L. M. (in preparation) Paving the way for implementing artificial multifunctional wetlands

Abstract

Wetland management requires a spectrum of scientific and socioeconomic input, especially within the framework of water purification and ecosystem development. Combining both ecosystem services into a single system is challenging, as detailed knowledge on and experience with this kind of integrated constructed wetlands is lacking. Therefore, information on the treatment performance of and the biotic interactions within wetlands is combined here to identify issues to be tackled prior to the implementation of multifunctional wetlands. On the one hand, pollutant reduction in natural treatment systems is highly variable and case-dependent, as illustrated by the removal efficiencies for BOD (50 – 90 %), nitrogen (14 – 86 %) and phosphorus (35 – 91 %). Further understanding on how processes are affected by environmental conditions and how discharges affect the receiving water body are crucial for wide-scale application. On the other hand, a variety of biotic interactions occurs within shallow water systems and illustrates the essential role of macrophytes towards habitat creation. Their steering role regarding wetland community structure and functioning affects the physical, chemical and biological level and suggests that macrophytes are a potential starting point for wetland restoration and construction. Inference of the preferred abiotic conditions by means of occurrence-based correlative habitat suitability models provides potential, though highly depends on the quality of the available data, while biotic interactions are even harder to predict. Hence, additional attention towards model development, data quality assurance and controlled experiments offer the opportunity to fill these knowledge gaps. Moreover, specific attention should be given to invasive alien species as they often possess functional traits that differ from native species and, when given an opportunity following land use alterations or climate change, can alter the composition and functioning of native communities. Despite these challenges, artificial treatment wetlands provide the opportunity to counteract the ongoing loss of wetlands and related ecosystem services.

2.1 Setting the scene

In previous chapter, the concept of Integrated Constructed Wetlands (ICW) indicated the potential of artificial wetlands to mimic the intrinsic multifunctionality of natural wetlands. It highlighted that such integrated artificial wetlands combine both pollutant removal and biodiversity improvement, thereby directly affecting the surrounding local environment. Moreover, the effects of wetland presence resonate through space (and time), as wetlands increase landscape heterogeneity and provide potential for nutrient retention and cycling at the regional scale (Comín *et al.*, 2001; Gopal, 2016). At the watershed scale, they act as water buffer zones and increase connectivity between green zones, while regulating climate at the global scale (Gopal, 2016; Jenkins *et al.*, 2010; Mitsch and Gosselink, 2000) (see Figure 2.1). Typically, these effects take place faster at a smaller scale, while being temporally lagged at the larger scale.

The benefits of wetland presence are not limited to supporting a variety of environmental processes and cycles, but extends to providing the potential to combat current ecological, economic and societal issues including the occurrence of algae blooms in eutrophic freshwater systems, the presence of dead zones near river mouths and coasts (Breitburg *et al.*, 2018), the salinisation of coastal and freshwater wetlands (Herbert *et al.*, 2015), the ongoing acidification of rivers and lakes (Weiss *et al.*, 2018), the increasing rainfall intensity and flooding frequency (Kundzewicz *et al.*, 2014), the depletion of groundwater (Döll *et al.*, 2014), global and personal human health issues (Hartig *et al.*, 2014) and the required development towards a more circular economy (Singh and Ordoñez, 2016), as exemplified by Figure 2.1.

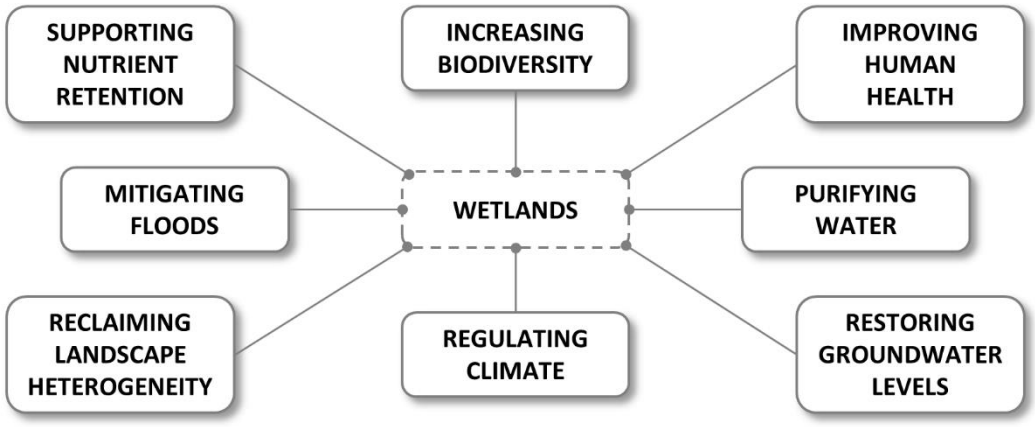

Figure 2.1: Examples of ecosystem processes and services provided by wetlands. *Several processes are beneficial towards both ecosystem functioning and society (e.g. purifying water, mitigating floods), while others are more specifically beneficial towards the environment (e.g. restoring groundwater levels) or society (e.g. improving human health).*

Optimising and safeguarding these ecosystem services requires a concert of scientific, societal and economic input that ultimately results in the protection, restoration or construction of wetlands (Jackson *et al.*, 1995; Kingsford *et al.*, 2016; Whigham, 1999). Protection requires the least input as it merely applies to (mostly) pristine systems with relatively high ecological and economical value. The prevailing conditions within these systems support a stable and undisturbed situation and a diverse biological community. Wetlands displaying a decrease in ecological quality (ongoing or historical) may benefit from human intervention, hence restoration is often applied. Yet, despite the availability of existing principles, restoration actions are carried out on a case-by-case basis with low repeatability (Keddy, 1999). Finally, wetland construction is performed to mimic natural systems and profit from the delivered ecosystem services. For instance, constructed treatment wetlands (CTWs) are highly tuned systems for the mere optimisation of pollutant removal. The majority of these systems consists of a herbaceous species growing in a substrate with wastewater flowing either on top (free water surface; FWS) or through (sub-surface flow; SSF) the substrate (Vymazal, 2010).

Within this chapter, specific attention is given to the construction of wetlands that allow (i) direct interaction with the atmosphere, (ii) the presence of both shallow and deep zones, (iii) the establishment of macrophytes and (iv) the creation of microhabitats via compartmentalisation. To avoid confusion with natural wetlands, terminology from the field of artificial treatment wetlands will be used further on, referring to the preferred wetland as a Free Water Surface (FWS) wetland (Gopal, 2016; Kadlec, 2009; Vymazal, 2010). The construction and restoration of these FWS CTWs provide a unique opportunity to create a single answer to two separate problems: (i) direct discharge of eutrophic wastewaters into the environment and (ii) loss of wetland-related biodiversity and the according ecosystem services. Therefore, an assessment is made throughout the following sections of both the chemical processes and biotic interactions occurring within the FWS system, with specific attention towards phytoplankton, periphyton, zooplankton, macroinvertebrates, macrophytes and fish. This contrasts with the main biotic focus brought forward in Section 1.2.1, but was considered essential when introducing ICWs.

The aim is to create an overview of how specific biotic groups interact in shallow eutrophic freshwater systems and, from that, derive which biotic group(s) can provide a biological basis for developing a complex community. By tackling these two objectives, an answer is provided to RQ1.1 from Chapter 1. In addition, information on the biotic interactions and chemical treatment performance is combined to identify key issues to improve the implementation of multifunctional artificial wetlands, thereby answering RQ1.2. Therefore, this chapter concludes with a summary of the identified key issues.

2.2 Pollutant removal within constructed treatment wetlands

Wetlands receive the majority of their resources from terrestrial systems and offer a useful combination of conditions for supporting the biogeochemical cycles locally (Keddy, 2010). For instance, due to their sink function, wetlands accumulate relatively high amounts of carbon (Dodds and Whiles, 2010), which is used as a food source by the prevailing microorganism community (see Box 2.1). Aside from creating new biomass, gaseous carbon-based by-products are excreted by these microorganisms, including CO_2 (under aerobic conditions) and CH_4 (under anaerobic conditions). Yet, due to the high amount of carbon within the wetland, oxygen is often depleted throughout the majority of the water column, causing mostly anaerobic conditions to occur. Consequently, wetlands tend to contribute to climate change by exhausting CH_4 and N_2O, which are created in anaerobic conditions and have a higher global warming potential than CO_2 (i.e. around 28 and 265 times at the century scale (IPCC, 2014), respectively). At the water surface, however, oxygen diffuses into the water column and allows for the presence of an aerobic boundary layer. Due to this layer, a heterogeneous environment exists, with complementary processes occurring in the top (aerobic) and bottom (anoxic) layers.

These biochemical processes have been the basis for applying a wetland configuration within the framework of water treatment, with wastewater originating from domestic, agricultural, industrial and storm water sources (Kadlec and Wallace, 2008; Vymazal, 2010). Similar to conventional treatment systems, these natural counterparts rely on the activity of microorganisms to mineralise or transform waste products into new resources (i.e. nutrients, see further) without extensively applying chemicals, electricity or artificial aeration, although research on how these factors impact treatment performance is ongoing (Donoso *et al.*, 2019; Gao *et al.*, 2017).

The microbial conversion of organic material into new resources supports the survival and reproduction of primary producers (phytoplankton, macrophytes). Moreover, due to their sink function and associated biogeochemical processes, freshwater wetlands can produce up to 10 times more biomass than lakes and streams (i.e. around 1100 $g \cdot m^{-2} \cdot y^{-1}$ versus 110 $g \cdot m^{-2} \cdot y^{-1}$, respectively) (Dodds and Whiles, 2010). This biomass, in turn, acts as a food source for heterotrophic organisms (zooplankton, macroinvertebrates, fish), including herbivores and detritivores. Hence, the presence of these biogeochemical processes provides the basis for complex food web development.

Within the remainder of this section, the attention is focused on (1) the most frequently occurring and reported pollutants within CTWs and (2) additional key aspects that require study to improve and evaluate the applicability of CTWs. The different biotic groups that benefit from the provided resources will be discussed in the next section (see Section 2.3).

Box 2.1: Microorganisms in constructed treatment wetlands

The terminology 'microorganism' as used here overarches several biotic groups, including Archaea, bacteria, fungi and microscopic algae. The identification and classification of microorganisms is increasing, though researchers acknowledge the idea that the majority remains undiscovered (Cavicchioli et al., 2019; Saccá et al., 2017). The composition of such a microorganism community is highly case-dependent and often hard to control due to the complex interplay and dependency among species. For instance, He *et al.* (2017) indicated that the increased usage of saltwater to replace freshwater during activated sludge treatment potentially affects the performance of the system and illustrated that increased salinity decreased bacterial activity and sludge floc size. Nevertheless, microorganism presence remains essential in developing and maintaining the biogeochemical nutrient cycles that underlie the high-valued attenuation capacity of natural systems (Cavicchioli et al., 2019; Saccá et al., 2017).

Within the considered FWS CTWs, microorganisms can occur in the sediment, in the sludge layer, suspended in the water column and attached to alternative substrates (including stones, vegetation, liners and pipes). The latter often combines with non-motile algae and the resulting micro-community is generally referred to as periphyton, which is described in more detail in Section 2.3.1.2, along with its importance for supporting the development of aquatic food webs. Transformation of pollutants throughout CTWs is highly dependent on the activity of these microorganisms and therefore benefits from the creation of additional surface area. Consequently, higher removal efficiencies are theoretically obtained for treatment systems characterised by a flow through a substrate (i.e. subsurface flow) rather than on top of a substrate (i.e. free water surface), although this has been contradicted by field observations (Kadlec, 2009).

Presence of microorganisms within CTWs is crucial for developing aquatic food webs, while a variety of factors (e.g. temperature, wastewater type, macrophyte presence) dynamically influences the prevailing community composition. For instance, Wang *et al.* (2016) observed that reduced temperatures negatively affected the performance of the microorganisms and, consequently, the removal efficiency of the system. Moreover, they found that plant presence has a positive effect on microbial abundance, being further extended by Hernández-Crespo *et al.* (2016) who stated that combining multiple plant species supports a more diverse microbial community.

2.2.1 Wastewater pollutants and removal within CTWs

Within this subsection, specific attention is given to total suspended solids (SS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen (tN) and total phosphorus (tP), as they are the main focus of both legislation and research. Complementary topics dealt with in literature include metals, pharmaceuticals and personal care products (PPCP), pesticides, faecal contamination and endocrine disruptors (ED) (Vymazal, 2009).

SS represent all the particulate matter being suspended in the water column, covering a fraction of the overall BOD, COD, tN and tP content. Due to the low flow conditions within FWS CTWs, SS is mostly reduced via settling and complemented with decantation and filtering (e.g. due to macrophyte presence) supporting removal efficiencies up to 80 % (Kadlec and Wallace, 2008; Rousseau *et al.*, 2004b; Verhoeven and Meuleman, 1999). Consequential to the settling of SS, a sludge layer is formed at the bottom, being a mix of non-degradable (e.g. sand, silt) and degradable solids, allowing the latter to dissociate and, ultimately, dissolve within the water column or dissipate into the atmosphere. The mineralisation underlying this dissociation is a complex concert of pollutant-specific processes (see further), often resulting in reduced sludge volumes.

Within both the settled solids and water column, organic pollutants (BOD and COD) are subjected to biochemical processes conducted by microbial activity. Both aerobic respiration (conversion of organic-C into CO_2) and anaerobic fermentation (conversion of organic-C into CH_4) support the (partial) removal of BOD and, hence, COD (Kadlec and Wallace, 2008). The openness of FWS allows for the diffusion of oxygen into the water column, yet this rate tends to be lower than the overall oxygen demand and causes oxygen depletion near the bottom. The resulting gradient separates the aerobic layers with CO_2-production at the surface from the anaerobic layers with CH_4-production near the bottom. Still, removal efficiencies up to 90 % for BOD and 80 % for COD have been reported, though these can be as low as 50 % for BOD and 60 % for COD (Galanopoulos *et al.*, 2013; Healy *et al.*, 2007; Kivaisi, 2001; Wang *et al.*, 2017).

At nutrient level, both sedimentation and microbial activity play a role, though the importance of each process depends on the type of nutrient under consideration. For instance, the nitrogen cycle is highly diverse, including anions (NO_2^- and NO_3^-), cations (NH_4^+) and gaseous forms (NH_3, N_2O and N_2). With N_2 being the main component of the atmosphere, the transformation of organically bound nitrogen via ammonification (production of NH_4^+ and NH_3 following a pH-based equilibrium), nitrification (conversion of NH_4^+ into NO_3^- via NO_2^- in aerobic conditions) and denitrification (conversion of NO_3^- into N_2 in anaerobic conditions with a C-source) into nitrogen gas (see Figure 2.2) does not pose any significant environmental impact.

However, the gaseous nitrous oxide (N_20) produced during incomplete denitrification potentially leaks into the atmosphere where it contributes to global warming (IPCC, 2014; Song *et al.*, 2012). Moreover, with nitrification occurring in the aerobic top layer and denitrification taking place in the anaerobic (sludge) layer (Figure 2.2) (Vymazal, 2010), the overall nitrogen removal efficiency is highly dependent on the diffusion process, resulting in lower values compared to carbon removal efficiencies (up to 86 %, but going as low as 14 % (Wang *et al.*, 2017)). Nevertheless, when all conditions are present to support ammonification, nitrification and denitrification, FWS CTW can remove nitrogen indefinitely (Zedler, 2003).

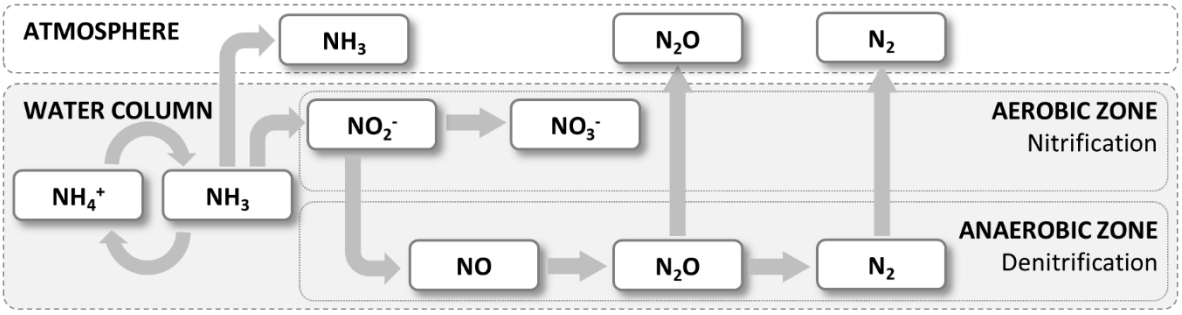

Figure 2.2: Illustration of the nitrogen cycle within wetlands. *Dissolved ammonium (NH_4^+) equilibrates with ammonia (NH_3), which is oxidised to nitrite (NO_2^-) and nitrate (NO_3^-) in the aerobic zone. In the anaerobic zone, nitrite is reduced to nitrous oxide (N_2O) and nitrogen gas (N_2). The latter two can escape into the atmosphere as a gas, as well as ammonia (NH_3).*

In contrast to carbon and nitrogen removal, limits occur with respect to phosphorus removal due to the absence of a gaseous form. Phosphate-ions (PO_4^{3-}) adsorb on the substrate surface, which, over time, results in lower removal efficiencies due to saturation effects (Bolton *et al.*, 2019; Vohla *et al.*, 2011). For instance, Wang *et al.* (2017) reported a decrease in phosphorus removal efficiency from 91 % down to only 35 % due to saturation effects within the substrate. Studies on characteristic substrate saturation curves indicate that phosphorus breakthrough in operational CTWs can be delayed by using different substrates (Bolton *et al.*, 2019; Park and Polprasert, 2008). Unfortunately, no stand-alone solutions to this substrate saturation are currently available, which implies that CTWs cannot act as completely independent treatment systems.

Each of these removal processes is steered by a plethora of abiotic variables, ranging from manageable (e.g. retention time, substrate depth) to unmanageable (e.g. temperature, precipitation) variables. Research related to these variables indicates an improved performance within a warmer climate, broad open spaces and higher retention times (Garfí *et al.*, 2012; Kadlec, 2009; Kotti *et al.*, 2010).

For instance, Wang *et al.* (2017) observed that cold climates had an upper limit of 80 % removal for BOD, while Kadlec and Wallace (2008) reported BOD removal efficiencies over 80 % in warm climates, thereby illustrating the positive effect of increased metabolic rates of the prevailing microorganism assemblage due to elevated temperatures (Kadlec, 2009). In contrast, the removal of SS is negatively correlated with temperature, as elevated microorganism productivity increases their suspension potential, hence causing higher SS levels to occur within the wetland effluent compared to the influent (Kadlec and Wallace, 2008).

Similarly, steering biotic variables range from manageable (e.g. presence of vegetation) to limitedly or completely unmanageable (e.g. microorganism assemblage). The effect of macrophyte presence on pollutant removal has been studied for decades and has been reported as one of the factors controlling temperature and nitrogen removal in wetlands (García-Lledó *et al.*, 2011; Vymazal, 2007; Vymazal, 2013; Wang *et al.*, 2017). Macrophytes influence pollutant removal at several levels. For instance, plant structures within the water column provide a surface for microorganisms to grow and interact, thereby supporting improved contact between the organic pollutants in the water phase and the heterotrophic bacteria (Brix, 1997; Fan *et al.*, 2016). Moreover, within the root zone, oxygen is released and results in micro-aeration of the substrate, thereby locally supporting the presence of aerobic bacteria active in the oxidation of both organic matter and nitrogen compounds (Brix, 1997; Vymazal, 2013). Contrasting these indirect effects, a direct effect on nutrients is exerted via uptake and assimilation into biomass (Beutel *et al.*, 2014; Dierberg *et al.*, 2002). However, this type of nutrient removal has been observed to account for maximally 10 % of the total incoming load and is potentially returned to the water phase when biomass is not harvested (Hernández-Crespo *et al.*, 2016; Merlin *et al.*, 2002; Park and Polprasert, 2008).

The extensive range of variables identified to exert an influence on wetland performance in combination with the reported case studies to be found throughout literature, illustrates that many research gaps still exist, especially due to the limited comparability of different systems (Thomaz and Cunha, 2010). Moreover, the effect of these variables is not restricted to altering pollutant removal, but extends to the water body receiving the effluent of the treatment system. Similar to the effect of conventional wastewater treatment plants (Ort and Siegrist, 2009; Zhou *et al.*, 2009), both quantity and quality of the effluent have the potential to cause changes in the abiotic conditions downstream of the CTW discharge point. However, the intensity of these changes remains highly dependent on the actual flow of the discharge, which is often several magnitudes smaller than conventional systems due to being applied at a smaller scale. Despite the importance of discharge flow, attention in the following section is mostly directed at the treatment performance of artificial wetlands.

2.2.2 Improving treatment to accommodate clean water and sanitation

Constructed treatment wetlands provide the potential to reduce the amount of incoming suspended solids and biodegradable organic compounds up to 85 % and 80 %, respectively, although this highly depends on the type of water being treated (Hijosa-Valsero *et al.*, 2010; Verhoeven and Meuleman, 1999). However, environmental conditions greatly affect microbial processes, causing difficulty in reaching stable effluent concentrations, while the absence of strong oxidative compounds within the treatment system impedes the removal of highly recalcitrant organic compounds (Donoso *et al.*, 2018). This provides two main areas for further research: (i) improve the understanding of how prevailing conditions affect the treatment efficiency and (ii) determine the potential impact of recalcitrant compounds on freshwater conditions.

Improved understanding of the treatment performance implies the combination of experiments, analyses and simulations. A multitude of experimental studies discussing separate case-studies can be found in literature, applying a range of wastewater compositions (Garfí *et al.*, 2012; Wang *et al.*, 2017), different kinds of vegetation (Maine *et al.*, 2007; Vymazal, 2013) and a variety of substrate types (Sakadevan and Bavor, 1998; Vohla *et al.*, 2011), yet provide a limited basis to support an overall, holistic comparison. For instance, Donoso *et al.* (2017) assessed the operating conditions (i.e. temperature, water flow) of FWS CTWs treating diffuse nutrient pollution and concluded that FWS CTWs provide an alternative measure to fight the eutrophication of waterways. Despite the fact that this result supports the applicability of FWS CTW as a mitigation measure, only superficial information related to the influence of prevailing conditions on treatment performance can be extracted from this type of studies. This highlights the need of more in-depth research to obtain a better process-based understanding of CTW performance and the inherent influence of environmental conditions.

Secondly, despite providing relatively high removal efficiencies for specific pollutants, trace concentrations do occur within effluents that are discharged into the environment, especially in the case of recalcitrant compounds. Effects caused by their discharge are highly case-specific and depend on the prevailing freshwater conditions on the one hand and on the pollutant load and discharge frequency on the other hand. For instance, exceeding the official effluent standards causes an unequivocal drop in absolute water quality, while the relative change can be higher for high-quality compared to low-quality surface waters. To illustrate this, Donoso *et al.* (2018) studied the relevance of COD discharge limits for CTWs treating animal manure by assessing the occurrence of macroinvertebrates in the receiving river. They observed the presence of pollution-sensitive taxa downstream of the discharge point, despite the standard-exceeding COD concentrations in the effluent, suggesting that the existing COD-standards might be too stringent.

Aside from indicating the limited environmental effect, Donoso *et al.* (2018) did not specify the COD compound composition, making this kind of conclusion inference overly simplistic and inappropriate towards other types of wastewater. For instance, high concentrations of insecticides can result in high COD concentrations in the effluent, simultaneously causing drastic effects on the downstream macroinvertebrate assemblage. Hence, a more in-depth characterisation of COD compounds and how they behave within the CTW is required prior to adapting the standards.

Progress within these fields is crucial to optimise the treatment process and limit the environmental impact. This requires the collective consideration of societal, environmental and operational aspects (Becerra Jurado *et al.*, 2009; Mereta *et al.*, 2012; Truu *et al.*, 2009), as illustrated in Figure 2.3. However, most studies only focus on a subset of these aspects, with limited research applying a holistic approach. For instance, De Troyer *et al.* (2016), assessed the water quality of the rivers and wetlands around Jimma (Ethiopia), considering both chemical and biological indicators. They acknowledged the potential of wetlands as a promising technique for wastewater treatment, though concluded that further societal awareness and stakeholder participation were needed to implement CTWs in regions affected by water pollution, limited sanitation and overall poverty. Similarly, other reports highlighted the capacity of natural and CTWs for wastewater treatment, while concluding that implementation is impeded due to stakeholders lacking insight into the integrated functioning of CTW ecosystems (Donoso *et al.*, 2017; Hefting *et al.*, 2013; Vymazal, 2010). These observations highlight the need for (i) including societal aspects into CTW research and (ii) assigning a budget for educating and involving local communities, confirming that restoration success is determined by merging science, society and politics (Catalano *et al.*, 2019; Jähnig *et al.*, 2011).

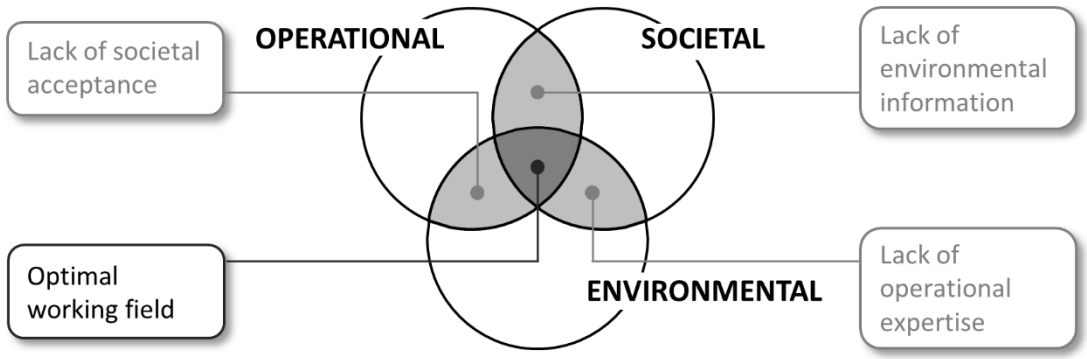

Figure 2.3: Illustration of the required input to improve implementation of constructed treatment wetlands. *By combining only two aspects, successful long term implementation is impeded due to the lack of societal, environmental or operational input.*

2.3 Biodiversity improvement by constructed wetlands

2.3.1 Occurrence of and interactions between key biotic groups

Wetlands are highly diverse and complex systems and support the survival of a variety of biotic groups. Here, only a selection of them is discussed as an in-depth discussion of each group separately goes beyond the scope of this chapter. More specifically, the aim of this section is to identify the biotic group that provides a relatively strong steering effect on the development of a complex biotic community, by focusing on their functioning within the trophic food chain and contribution to creating a concert of microhabitats. Therefore, two primary producers are considered (phytoplankton and macrophytes) along with grazers (zooplankton, macroinvertebrates and fish) and aquatic predators (macroinvertebrates and fish), supplemented with a mixed group of autotrophic and heterotrophic organisms (periphyton). Despite the interactions displayed by amphibians, mammals, bats and birds as important energy linkages with the terrestrial system (Chawaka *et al.*, 2018; Gopal, 2016; Parker *et al.*, 2019), they are not discussed here. An overview of the selected biotic groups and additional information can be found in Table 2.1 and subsequent sections.

Table 2.1: Glossary for the biotic groups discussed within this chapter. *For each biotic group (phytoplankton, periphyton, zooplankton, macroinvertebrates, macrophytes and fish) a short description and main subgroups are provided. Their importance within shallow, eutrophic systems is further specified throughout the group-specific subsections.*

Biotic group	Description
Phytoplankton	Free-floating group of microscopic organisms containing chlorophyll to capture sunlight, with most important subgroups being the cyanobacteria, green algae, diatoms and dinoflagellates. Within shallow freshwater systems, Bacillariophyceae, Chlorophyceae, Cyanophyceae and Euglenophyceae are frequently reported (Calero *et al.*, 2015; Chen *et al.*, 2011b; Travaini-Lima *et al.*, 2016; Vincent and Kirkwood, 2014), with varying community composition depending on both climatic and operating conditions. Their growth is supported by sunlight, carbon dioxide (CO_2) and nutrients.
Periphyton	Group of microscopic organisms consisting of green algae, cyanobacteria and (heterotrophic) microorganisms. They mostly occur in symbiosis attached to submerged surfaces, including substrate, vegetation and non-natural constructions. Their growth is supported by the interaction between the autotrophic (sunlight, carbon dioxide and nutrients) and heterotrophic (organic compounds and by-products) species.

(Continues on next page)

(Continued)

Biotic group	Description
Zooplankton	Actively moving organisms that lack chlorophyll to provide in their energy requirements, hence their heterotrophic feeding behaviour. The most important groups to be considered within shallow freshwater systems are cladocerans, rotifers, copepods and ostracods. Their growth is mainly supported by the presence of phytoplankton, detritus and other zooplankton species. Several species belong to the Crustacea and are key primary consumers in lotic systems (Dodds and Whiles, 2010).
Macroinvertebrates	Macroinvertebrates are small organisms without a backbone, but large enough to be seen with the naked eye. They mostly live in the benthic layer, but species living near the water surface and within the water column exist as well. They feed on detritus, plankton (both suspended and settled), other invertebrates and plants. Macroinvertebrate monitoring is a common technique to assess the biological water quality as this group is rather diverse and ranges from pollution-sensitive to pollution-tolerant taxa, making them ideal surrogates for assessing wetland health (Balcombe *et al.*, 2005a).
Macrophytes	Macrophytes represent all types of aquatic vegetation that can be found within a shallow water body and in the littoral zones of rivers, lakes and oceans. A distinction is made between nonvascular (e.g. mosses, known as bryophytes) and vascular (e.g. reed, duckweed) plants, of which the latter is often subdivided in emergent, submerged and floating plants (Dodds and Whiles, 2010). Macrophytes require nutrients, carbon dioxide and sunlight to create new biomass, hence a vast amount of research on their applicability as pollutant removers (i.e. phytoremediation) has been performed (Brisson and Chazarenc, 2009; Hernández-Crespo et al., 2016; Rodríguez and Brisson, 2015; Tanner, 1996).
Fish	Highly diverse group with more than 10 000 freshwater species, feeding on a variety of food sources, ranging from phytoplankton over macroinvertebrates and macrophytes to other fish (Batzer et al., 2000; Dodds and Whiles, 2010). The most common freshwater fish orders (> 2000 species) are Cypriniformes, Siluriformes and Perciformes (Dodds and Whiles, 2010).

2.3.1.1 Phytoplankton

Considering the low flow and prevailing eutrophic conditions, FWS CTW provide an optimal environment for phytoplankton to grow and prosper, especially when the hydraulic retention time is high and macrophyte cover is limited (Luyiga and Kiwanuka, 2003). Within these systems, phytoplankton communities often indicate a dependence on system design, climatic conditions and nutrient concentrations. For instance, Travaini-Lima *et al.* (2016) associated the observed increase in biomass of *Kirchneriella lunaris* (class Chlorophycaceae) during the dry season with elevated nutrient levels entering the system. Similarly, Chen *et al.* (2011b) linked the difference in phytoplankton community between three different CTWs treating domestic wastewater with the prevailing total phosphorus concentration.

The value of phytoplankton within shallow freshwater systems is ambiguous and highly dependent on abundance (Zimmer *et al.*, 2003). For instance, at low concentrations, they mainly take up nutrients and carbon dioxide to create new biomass through photosynthesis, thereby positively supporting the development of higher trophic levels. In contrast, exudates originating as by-products from metabolic processes can decrease flocculation and subsequent settling of suspended solids, thereby negatively affecting transparency and, thus, wetland treatment performance (Sun *et al.*, 2013).

At high concentrations, algae blooms can develop due to the uncontrolled proliferation in eutrophic conditions, which can lead to fluctuating oxygen levels that reach complete absence of oxygen. Limitation of oxygen supports the production and volatilisation of ammonia and negatively influences organisms that rely on respiration for their energy balance (e.g. macroinvertebrates, fish), which ultimately limits their survival (Luyiga and Kiwanuka, 2003; Miranda and Hodges, 2000). Moreover, some species (especially cyanobacteria) excrete toxic compounds threatening fish population and human health (Dodds and Whiles, 2010; Vincent and Kirkwood, 2014), requiring a bottom-up (nutrient control) or top-down (biological or chemical control) approach.

More specifically, macrophytes compete with phytoplankton for nutrients and limit the amount of light entering the water, hence limiting the presence of algae (Travaini-Lima *et al.*, 2016; Zimmer *et al.*, 2003). Simultaneously, the excretion of allelochemicals (e.g. phenolic compounds) can inhibit algae growth, although this highly depends on the specific macrophyte-algae interaction (Zhong *et al.*, 2016). In contrast to this resource limitation, zooplankton and fish exert a top-down control strategy as they feed on phytoplankton (Fontanarrosa *et al.*, 2010). For example, Calero *et al.* (2015) observed an increase in zooplankton biomass up to 64 % in the Albufera Lake FWS CTW along a decrease of phytoplankton biomass, suggesting the presence of a trophic interaction.

2.3.1.2 Periphyton

Due to the artificial nature of CTWs, surface areas for colonisation by periphyton can be optimised to improve the contact between pollutants and bacteria within the treatment system (Gao *et al.*, 2019). Simultaneously, the reduced flow conditions support the settlement of suspended solids and, hence, the penetration of light through the water column. Improved light conditions benefit the development of algae within the periphyton layer where bacterial mineralisation provides additional resources to create algal biomass (Ishida *et al.*, 2008; Sand-Jensen and Borum, 1991; Toet *et al.*, 2003).

Periphyton in natural systems consists of a complex microbial community and can be characterised by a highly dynamic species turnover without major effects on the overall periphyton functioning (Liu *et al.*, 2016). Due to this complex composition and dynamics of the periphyton layer, most studies report on treatment efficiency and ignore or only partially describe the species composition (Rooney *et al.*, 2020; Zamorano *et al.*, 2018). Yet, Cronk and Mitsch (1994) analysed the periphyton composition of four wetlands under different hydrologic conditions and observed Bacillariophycaceae (4 genera), Chlorophyta (6 genera) and Cyanophyta (1 genus), though acknowledged that the system might have been too immature to support a developed periphyton community.

Presence of periphyton is controlled by both abiotic and biotic conditions. The composition of the periphyton layer depends on (1) the presence of degradable organic compounds (mineralisation by bacteria) and (2) the presence of absorbable nutrients (photosynthesis by algae). The relative presence of these resources contributes to the final periphyton composition (Wu *et al.*, 2018). At the biotic level, both indirect and direct interactions occur and reflect a certain degree of similarity with phytoplankton. Competition of periphytic algae with phytoplankton and macrophytes for nutrients can occur (Sand-Jensen and Borum, 1991), though is countered by the symbiosis with mineralising bacteria in the vicinity (Liu *et al.*, 2017).

In contrast, shading by phytoplankton and macrophytes has a clear negative effect on light availability and, thus, on the development of algae within the periphyton layer (Sand-Jensen and Borum, 1991; Toet *et al.*, 2003). In addition, periphyton is exposed to grazing by organisms from higher trophic levels. A variety of zooplankton, macroinvertebrate and fish species rely on the presence of periphyton to provide in their nutritional needs (Batzer and Resh, 1991; Rooney *et al.*, 2020; Sand-Jensen and Borum, 1991).

Aside from providing a positive contribution to the overall pollutant reduction in the treatment system, additional support for the development of macrophytes can be provided. Macrophytes covered with periphyton can benefit from the locally produced nutrients instead of relying on the diffusion of nutrients within the water column (Gao *et al.*, 2019).

2.3.1.3 Zooplankton

Due to the low-flow conditions and potential high phytoplankton presence (see above), FWS CTWs act as nurseries for zooplankton with biomass increasing throughout the system, especially when macrophytes are present (Calero *et al.*, 2015; Hernández-Crespo *et al.*, 2017). The zooplankton community is frequently dominated by cladocerans or rotifers and exceeds diversity in drains and rivers (Eivers *et al.*, 2017), with sporadic seasonal variation in community composition (Beaver *et al.*, 1998; Calero *et al.*, 2015; Travaini-Lima *et al.*, 2016). For instance, Travaini-Lima *et al.* (2016) observed that rotifers dominated in both the rainy and dry season, with overall higher zooplankton density during the rainy season. Similarly, Calero *et al.* (2015) found clear seasonal fluctuations in zooplankton biomass, with rotifer dominance in summer, copepod dominance in winter and cladocerans dominating in spring.

At the biotic level, zooplankton is mainly influenced by phytoplankton, fish and macrophytes, either directly or indirectly (Table 2.2). The interactions with phytoplankton and fish represent a straightforward trophic cascade interaction, with zooplankton feeding on phytoplankton and fish consuming zooplankton (Calero *et al.*, 2015; Cao *et al.*, 2007). More importantly, the selective preying by fish causes shifts in zooplankton communities and has a tendency of altering the male-to-female ratio (thus affecting the associated population dynamics) (Bramm *et al.*, 2009).

Macrophytes act supportively as a refuge area for zooplankton to escape from fish predation (diel horizontal migration, DHM) and provide a habitat for cladoceran diapausing eggs (Calero *et al.*, 2015; Castro-Castellon *et al.*, 2016; Travaini-Lima *et al.*, 2016). Yet, despite the creation of physical habitats, macrophytes negatively affect light conditions (e.g. dense duckweed mats) and thereby reduce the quality and quantity of the zooplankton community, resulting in a lower zooplankton diversity compared to high light conditions (Bramm *et al.*, 2009; Fontanarrosa *et al.*, 2010). Moreover, when planktivorous fish abundance is high, predation pressure increases and DHM becomes limited (Meerhoff *et al.*, 2007).

Still, high zooplankton densities are not necessarily linked with high phytoplankton densities. For instance, Kampf and Claassen (2004) observed high zooplankton densities while phytoplankton was almost absent and inferred that zooplankton also survived by consuming bacteria. As such, they suggested to culture *Daphnia magna* with treatment plant effluents prior to their use as food source for sticklebacks (Kampf and Claassen, 2004).

2.3.1.4 Macroinvertebrates

Despite the relatively high pollutant levels, specific macroinvertebrates are able to survive within FWS CTWs due to the presence of adequate food sources (Becerra Jurado *et al.*, 2009; Boets *et al.*, 2011; Céréghino *et al.*, 2008; Chen *et al.*, 2011b; Hsu *et al.*, 2011). Recurring observations in natural and artificial wetlands include Coleoptera and Hemiptera as dominating orders and the influences of season and wetland conditions on macroinvertebrate community composition (Becerra Jurado *et al.*, 2009; Boets *et al.*, 2011; Céréghino *et al.*, 2008; Fairchild *et al.*, 2000).

For instance, Becerra Jurado *et al.* (2009) found 123 taxa in 15 constructed wetlands treating wastewater, dominated by Coleoptera (45 %) and Hemiptera (17 %), though did not provide a detailed study on the influence of season. In contrast, Boets *et al.* (2011) investigated a single FWS CTW in summer and autumn and reported a higher taxa diversity in summer dominated by Corixidae (Hemiptera) and Chironomidae (Diptera), next to an overall increase in diversity along the treatment path (representing a decrease in nutrient levels). Additionally, Robson and Clay (2005) observed that seasonal wetlands had less taxa than perennial wetlands due to higher levels of temporal variation, although both could still be considered as taxon-rich.

Macroinvertebrates experience direct and indirect influences, originating from zooplankton, fish, macrophytes and even higher-order animals (Table 2.2), though indicate to be highly taxon-specific. For instance, Corixidae and Veliidae (Hemiptera) benefit from fish presence, while being part of the diet of dabbling ducks (Balcombe *et al.*, 2005a). Similarly, Planorbidae (Mollusca) benefit from the presence of macrophytes because of their grazing activity, but can be suppressed by predatory fish, which results in a simultaneous positive effect on epiphytic chironomid larvae (Batzer *et al.*, 2000). Still, macroinvertebrates provide several useful functions within wetlands, ranging from litter decomposition over plant community regulation to nutrient cycling towards higher trophic levels (including waterfowl and anurans), due to their place in the food chain and the potential of several insects to switch from an aquatic to a terrestrial stage in their life cycle (Balcombe *et al.*, 2005a; Dodds and Whiles, 2010; Hsu *et al.*, 2011; Knight *et al.*, 2001).

Wetlands are said to be easily colonised by macroinvertebrates, requiring about four to five years to reach maximal species diversity (Hansson *et al.*, 2005). This can be facilitated by proximity of other ponds (i.e. high connectivity) (Céréghino *et al.*, 2008; Nelson *et al.*, 2000), although Robson and Clay (2005) did not observe a specific species assemblage of closely located sites. Most importantly, macroinvertebrates within these FWS CTWs are highly system-specific due to the unique prevailing abiotic conditions and thereby contribute to the overall catchment diversity (Becerra Jurado *et al.*, 2009; Céréghino *et al.*, 2008).

2.3.1.5 Macrophytes

Nutrient concentrations within FWS CTWs are sufficiently high for macrophytes to grow, with presences reported in a variety of wetland types, ranging from small-scale domestic wastewater treatment systems over floating wetlands to large-scale restoration wetlands (Castro-Castellon *et al.*, 2016). Vegetation is often emergent, including common reed (*Phragmites australis*), cattail (*Typha latifolia* and *T. angustifolia*), sedge (*Carex acutiformis*) and bulrush (*Schoenoplectus* spp.) (Brisson and Chazarenc, 2009; Castro-Castellon *et al.*, 2016; Rodríguez and Brisson, 2015), though also floating plants have been reported, including water hyacinth (*Eichhornia crassipes*), water lettuce (*Pistia stratiotes*) and duckweed (*Lemna* spp.) (Hsu *et al.*, 2011; O'Farrell *et al.*, 2009).

Observed effects, including microaeration of the root zone, provision of substrate for periphyton development and limiting sediment resuspension, suggest that certain macrophyte species are effective ecosystem engineers within shallow wetland systems (Brix, 1997; Gopal, 2016; Vymazal, 2011b). For instance, a higher diversity of macroinvertebrate taxa was observed in vegetated areas compared with non-vegetated areas, due to a decreased risk of predation, a complex spatial structure and being a location for cladoceran diapausing eggs (Stiers *et al.*, 2011; Timms and Moss, 1984). Moreover, also waterfowl benefit from the presence of emergent macrophytes for nesting and roosting, being at the same time close to an appropriate food source (Gopal, 2016).

Next to exerting a variety of influences on fish, macroinvertebrates, zooplankton and phytoplankton, macrophyte presence is prone to grazing (fish and macroinvertebrates) and competition for nutrients (phytoplankton) (Table 2.2). Grazing pressure remains limited due to the low total number of strictly herbivorous fish and macroinvertebrates. In contrast, competition with phytoplankton under eutrophic conditions can lead to complete disappearance of vegetation within a wetland by rapidly changing nutrient availability, light penetration and pH level (Lu *et al.*, 2012; Scheffer *et al.*, 1993a).

Presence of macrophytes can also have a negative effect on both chemical and biological conditions. For instance, dense vegetation stands decrease light penetration and oxygen concentrations (degradation of dead organic matter), thereby limiting respiration of higher trophic animals (Balcombe *et al.*, 2005a; Miranda and Hodges, 2000). However, Frodge *et al.* (1990) observed extremely high oxygen concentrations within the near-surface canopy of submerged macrophytes (going up to 30 $mg \cdot L^{-1}$), which dropped drastically when entering the sub-canopy zones (down to 1 $mg \cdot L^{-1}$ within 0.5 m). Hence, the creation of open water sections allows for species to migrate when needed, for phytoplankton to produce oxygen and fish to escape anoxia (Balcombe *et al.*, 2005a; Miranda and Hodges, 2000).

2.3.1.6 Fish

Presence of fish within FWS CTWs is only limitedly reported and if so, abundances are low (Chen *et al.*, 2011b; Hansson *et al.*, 2005; Hsu *et al.*, 2011; Kampf and Claassen, 2004). For instance, Kampf and Claassen (2004) pointed out that, although food was abundantly present in the FWS, no fish were observed, potentially due to high ammonia (NH_3) concentrations caused by exceeding the nitrification capacity of the treatment plant. Additionally, anoxic conditions, low winter and high summer temperatures and limited refuge areas represent a harsh environment for fish (Batzer *et al.*, 2000). However, when hydraulic retention time (HRT) became higher than two days, fish were observed as overloading was reduced (Kampf and Claassen, 2004).

Fish primarily provide top-down control on phytoplankton, zooplankton and macroinvertebrates (see above and Table 2.2), but are only limitedly influenced by these (leaving food availability aside). For instance, dense stands of both phytoplankton and macrophytes can lead to diel fluctuations in oxygen concentration and pH, representing unfavourable conditions for fish (Hsu *et al.*, 2011; Miranda and Hodges, 2000).

Yet, negative effects of fish presence have also been observed towards amphibians, with salamanders and tadpoles being frequently consumed by fish, sometimes even causing rapid extinction of the amphibian community after colonisation (Alford and Richards, 1999; Dodds and Whiles, 2010). Amphibians represent an important link between the aquatic and terrestrial environment, providing an alternative pathway for nutrient removal and a valuable link in complex food webs (Balcombe *et al.*, 2005b; Davic and Welsh, 2004). However, elevated nutrient and ion concentrations occurring within treatment wetlands limit the potential of amphibian presence and suggest that increased connectivity of the wetland with surrounding freshwater bodies might be more appropriate to increase overall diversity and nutrient transport via fish migration (Becerra-Jurado *et al.*, 2012; Wiegleb *et al.*, 2017).

2.3.1.7 Overview of biotic interactions

Table 2.2: Non-exhaustive overview of interactions among biological elements as reported in literature dealing with eutrophic, shallow water bodies. *Interactions describe the effect of the biotic group on a specific row on a biotic group in a specific column. PhP: Phytoplankton, PeP: Periphyton, ZP: Zooplankton, MI: Macroinvertebrates, MP: Macrophytes.*

	Phytoplankton	Periphyton	Zooplankton	Macroinvertebrates	Macrophytes	Fish
PhP	- Cyanobacteria can produce toxins[11] - Self-shading[11, 21]	- Cyanobacteria can produce toxins[11] - Light interception due to blooms - Competition for nutrients	- Serve as food source[9, 12, 20]	- Cyanobacteria can produce toxins[11] - Serve as food source[3, 11] - Anoxia due to algae blooms	- Cyanobacteria can produce toxins[11] - Light interception due to blooms[23, 25] - Competition for nutrients[27]	- Cyanobacteria can produce toxins[11] - Increased turbidity[28] - Diel fluctuations in oxygen and pH[1] - Anoxia influences growth, swimming speed and survival[18]
PeP	- Competition for nutrients[19]	- Competition for nutrients	- Serve as food source	- Serve as food source[16, 19]	- Light interception due to blooms[23, 25] - Competition for nutrients[19, 27] - Provision of nutrients[13]	- Serve as food source
ZP	- Provide top-down control via grazing[9, 12]	- Provide top-down control via grazing	- Competition for the same food source[17]	- Serve as food source[17]	- Indirectly reducing the competition with phytoplankton	- Serve as food source[5, 17, 29]

(Continues on next page)

(Continued)

	Phytoplankton	Periphyton	Zooplankton	Macroinvertebrates	Macrophytes	Fish
MI	- Grazing[3] - Production of CO_2 and release of nutrients[26]	- Grazing[3, 8, 16, 19] - Production of CO_2 and release of nutrients	- Grazing[3]	- Competition for same food source[2, 3, 4] - Predation[2, 3, 4] - Shredders facilitate collectors by excreting fine organic matter[11]	- Herbivory[8, 11]	- Serve as food source[4, 14, 16]
MP	- Intercept light, shading of water, impeding algae growth[12, 24] - Competition for nutrients[25, 29] - Allelochemicals with negative, neutral or positive effect on phytoplankton growth[25, 27]	- Provide substrate to grow on[6] - Intercept light, shading of water, impeding algae growth[12, 24] - Competition for nutrients[19, 25, 29] - Indirectly affect grazing pressure by macroinvertebrates[3]	- Attached biofilm as food source[3] - Refuge sites in case of low fish density[9, 17, 23, 24, 25] - Dense mats can limit light and oxygen[1] - Exudates can have influence on migration[7, 25] - Support higher densities[10] - Habitat for cladoceran diapausing eggs[9]	- Direct food source (dead & alive)[2] - Indirect food source: attached biofilm[3, 22] - Refuge sites (e.g., midges sheltering from fish)[4, 14] - Habitat creation[4, 22, 25] - Influence on foraging efficiency[22] - Density influences community[3] - Oxygen source in anoxic environments[1] - Degradation can cause oxygen depletion[1]	- Competition[15] - Excretion of allelochemicals[15]	- Refuge area[11, 12, 22] - Habitat for egg deposition, larvae and juveniles[22, 25] - Light limitation can decrease foraging activity[5] - Direct food source[4] - Attracting prey[22] - Complexity influences visual contact with prey, foraging activity and growth[22] - Can cause diel patterns of pH and DO[1, 14, 18]

(Continues on next page)

(Continued)

	Phytoplankton	Periphyton	Zooplankton	Macroinvertebrates	Macrophytes	Fish
Fish	- Resuspension causes nutrient release from substrate[20, 26] - Excretion of nutrients[26] - Indirectly reducing predation pressure by zooplankton[29]	- Resuspension causes nutrient release from substrate[20] - Excretion of nutrients[26] - Grazing - Indirectly by predating on other grazers[16]	- Selective preying affects community composition (e.g. less crustaceans, more rotifers)[5] - Selective preying affects life history (e.g. higher male-to-female ratio of cyclopoids)[5, 25] - Chemical cues steer morphology and reproduction[7] - Diel migration[23]	- Feed on invertebrates, e.g. Chironomidae, Planorbidae, Physidae, Corixidae, Glossiphoniidae[4, 14, 16] - Indirect supporting macroinvertebrate presence via feeding on competitors or predators[2, 4]	- Herbivory[4] - Resuspension can limit submerse vegetation[14, 20, 29] - Excretion of nutrients[11]	- Competition for same food source (e.g. midges)[4] - Predation[4]

[1] Angélibert *et al.* (2004); [2] Balcombe et al. (2005a); [3] Batzer and Resh (1991); [4] Batzer et al. (2000); [5] Bramm et al. (2009); [6] Brix (1997); [7] Burks et al. (2000); [8] Carlsson and Brönmark (2006); [9] Calero *et al.* (2015); [10] Choi et al. (2014); [11] Dodds and Whiles (2010); [12] Fontanarrosa et al. (2010); [13] Gao *et al.* (2019); [14] Hsu et al. (2011); [15] Jarchow and Cook (2009); [16] Liboriussen *et al.* (2005); [17] Meerhoff et al. (2007); [18] Miranda and Hodges (2000); [19] Sand-Jensen and Borum (1991); [20] Schrage and Downing (2004); [21] Spieles and Mitsch (2000); [22] Thomaz and Cunha (2010); [23] Timms and Moss (1984); [24] Travaini-Lima et al. (2016); [25] van Donk and van de Bund (2002); [26] Vanni (2002); [27] Zhong et al. (2016); [28] Zimmer *et al.* (2000); [29] Zimmer et al. (2003)

2.3.2 Use of macrophytes to improve biodiversity

Macrophytes showed to provide a steering role regarding wetland community structure and functioning, affecting the physical, chemical and biological level. At the physical level, the presence of macrophytes reduces flow velocity and positively affects nutrient cycling and water storage. Moreover, in combination with macrophyte rooting, these reduced flow velocities cause less erosion and sediment resuspension, which positively affects transparency (Brix, 1997). However, under improved settling and decreased erosion, wetlands tend to be exposed to siltation and accretion, which can be further exacerbated by high transpiration rates of dense emergent communities (Angélibert *et al.*, 2004; Zedler and Kercher, 2004).

The consequences of these physical changes on wetland community composition and functioning are case-dependent and situated along the positive-negative continuum. For instance, Rooth *et al.* (2003) showed that the invasion of wetlands occupied by *Typha* spp. and *Panicum virgatum* in the Chesapeake Bay by the invasive *Phragmites australis* caused higher sediment accretion rates within the areas invaded by *P. australis*. Simultaneously, a reduction in total wetland area had occurred due to rising sea levels, yet the accretion caused by *P. australis* supported the continued existence of the invaded wetland. Hence, the invasion by *P. australis* caused the local disappearance of the native vegetation but avoided the complete loss of the wetland's functionality.

Secondly, at the chemical level, nutrients are taken up directly from the water column, the sediment or a combination of both. This uptake supports biomass production, carbon sequestration and phytoremediation (see Box 2.2), with the latter being of main research interest for several decades (Brisson and Chazarenc, 2009; Rodríguez and Brisson, 2015; Tanner, 1996). Yet, this direct nutrient removal is estimated to represent maximally 10 % of the total provided load, though can be increased when frequent harvesting is applied and biomass-incorporated nutrients are completely removed from the aquatic system (Merlin *et al.*, 2002; Park and Polprasert, 2008).

Within wetlands, oxygen is crucial for aerobic degradation and nitrification to occur (see Section 2.2.1). Emergent plants are known for providing root zone aeration within the (mostly anoxic) substrate, while being countered by an upward movement of methane (Bergström *et al.*, 2007; Keddy, 2010; Vymazal, 2011b). This oxygen provision oxidises the reduced nitrogen compounds and drives a continuous diffusion of both reduced and oxidised nitrogen by altering the equilibrium between substrate and water column concentrations (Keddy, 2010). However, extensive surface coverage and dead plant material entering the water column cause additional oxygen consumption and the release of immobilised nutrients. For instance, duckweed species (*Lemna* spp.) can form dense mats under eutrophic conditions, which causes relatively high mortality rates and associated oxygen depletion underneath the mats (Janse and Van Puijenbroek, 1998).

Box 2.2: Macrophytes providing phytoremediation to reduce pollutant levels

Throughout the past decades, macrophytes have been frequently applied to counter the presence of pollutants in contaminated soil or water (i.e. phytoremediation) (Arthur *et al.*, 2005; Dhir *et al.*, 2009). Processes vary from degradation over immobilisation to extraction and are highly species- and environment-specific. For instance, Zhao *et al.* (2015a) studied the potential of several floating duckweed species to recover nutrients from wastewater and observed that *Lemna japonica* provided the highest nitrogen and phosphorus recovery and removal rates, while producing the most protein-rich biomass. In contrast, Amon *et al.* (2007) investigated the ability of various emergent macrophytes in supporting the dechlorination and mineralisation of perchloroethylene and observed significant improvements in pollutant removal. Additional examples of phytoremediation being facilitated by aquatic macrophytes can be found in Carvalho *et al.* (2014), Dhir *et al.* (2009) and Rai (2009).

These effects of macrophyte presence on the physical and chemical conditions illustrate how species interact with their environment and create a framework for the development of biotic interactions (Vitousek *et al.*, 1997). For instance, the development of a stable and biologically complex ecosystem is highly dependent on the presence of food, preferably provided by (a community of) primary producers, as autotrophic biomass production acts as a basis for the trophic cascade, feeding zooplankton, macroinvertebrates, amphibians, fish and birds (Balcombe *et al.*, 2005b; Thomaz and Cunha, 2010; Worrall *et al.*, 1997).

Moreover, during this primary production, nutrients are continuously taken up from the surrounding environment and converted into organic compounds to support cell growth. This causes pollutant levels to decrease towards the downstream sections of vegetated treatment systems, which creates different abiotic habitats along the flow path (Caraco and Cole, 2002). Due to these decreasing pollutant levels, the biotic diversity has the potential to increase towards the discharge point as the prevailing pollutant levels are less restrictive (Becerra-Jurado *et al.*, 2012; Boets *et al.*, 2011).

From this, it is clear that macrophyte occurrence represents an interesting starting point to support the conservation of wetlands, despite being determined by a range of species-specific preferences, interactions and functional traits, including the abiotic environment, dispersal capacity, temporary tolerance, resource competition, population dynamics, community ecology and evolution (Guisan and Thuiller, 2005; Pulliam, 2000; Sinclair *et al.*, 2010). Appropriate wetland management requires that these aspects are considered into detail, with additional attention towards acceptable abiotic conditions for macrophyte presence.

2.3.2.1 Abiotic conditions for macrophyte presence

Field observations, laboratory experiments and expert knowledge contribute to an increased understanding of the preferred abiotic conditions (Hofstra *et al.*, 2020). Based on field observations, suitable abiotic habitats for macrophyte presence can be derived, reflecting the niche concept as postulated by Hutchinson (1957) and re-evaluated by Pulliam (2000). This niche is a *n*-dimensional hypervolume in which every point represents an environmental condition that supports indefinite species survival and is generally referred to as the *realised niche*. The *fundamental niche* extends the realised niche as it excludes the effects of biotic interactions like resource competition and predator-prey interactions, thereby merely reflecting the suitable abiotic conditions (Pulliam, 2000).

Despite being unfit for inferring the fundamental niche, observations are often used within a data mining environment to derive suitable habitats, predict species distributions, define conservation value and restrict the spread of invasive alien species (Araújo and Guisan, 2006; Elith and Leathwick, 2009; McPherson *et al.*, 2004). Information obtained through these modelling exercises provides a valuable contribution to the delineation of a species' realised niche, which allows its subsequent application as an overall filter, combining both abiotic and biotic influences (Anderson and Raza, 2010; Guisan and Rahbek, 2011). In contrast, experiments under controlled conditions allow to infer realistic species traits and population parameters, thereby aiding the development of process-based models with a more profound grounding in ecological theory (Gallien *et al.*, 2010). Due to this approach, process-based models are better positioned than data-driven models when aiming to delineate the fundamental (abiotic) niche (Kearney and Porter, 2009).

Modelling techniques aiming to delineate species niches are intrinsically situated along a continuum between purely data-driven and completely knowledge-based (Dormann *et al.*, 2012; Mount *et al.*, 2016; Van Echelpoel *et al.*, 2015), with observation-based habitat suitability models (HSMs) being highly data-dependent. Model performance and reliability rely on a plethora of variables, including the quality of the data and the applied model parameter settings, both of which require attention during model development (Everaert *et al.*, 2016; Marvin and John, 2003; Zhang *et al.*, 2003). Yet, despite their added value for ecosystem management, HSMs have been widely criticised in literature for a variety of reasons, including the limited consideration of species dispersal within the final model structure (Elith and Leathwick, 2009; Guisan and Thuiller, 2005; Jarnevich *et al.*, 2015). It is highly recommended to acknowledge these criticisms when assessing the applicability of modelling techniques.

2.3.2.2 Biotic interactions

The inclusion of biotic interactions builds further on the abovementioned abiotic preferences and can be considered as an additional filter that determines which species can occur, conditional to the prevailing community (Guisan and Rahbek, 2011). Theoretically, a variety of interactions can take place, including out-competition (disappearance of a species), restricting competition via exclusion (separate range), neutral interaction (shared range), facilitation (unidirectional range extension) and mutualism (bidirectional range extension), as illustrated in Figure 2.4.

These interactions occur mostly between macrophyte species, though additionally tend to cross the taxonomic boundaries between biotic groups, e.g. pollination, herbivory and parasitism (Guisan and Thuiller, 2005; Hofstra *et al.*, 2020). Moreover, due to these interactions, the fundamental niche approaches the realised niche and shows a decrease or enlargement of the tolerated and preferred abiotic conditions. More specifically, the underlying functional traits (e.g. biomass production, flowering, root:shoot ratio) are affected in a positive, neutral or negative way (see Box 2.3), with intensity and direction varying along the environmental gradients (Huntley *et al.*, 2004).

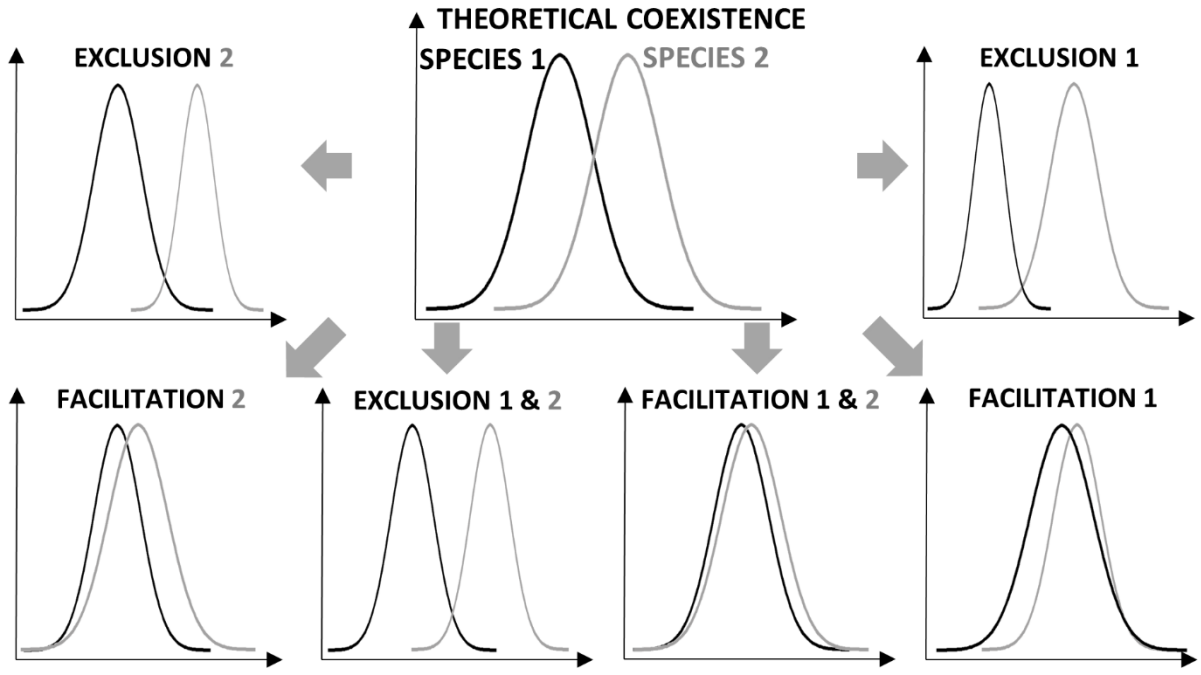

Figure 2.4: Illustration of the potential outcomes following biotic interactions between two species. *Each species is characterised by an occurrence frequency distribution (y-axis) over an environmental gradient (x-axis), which overlap when considered separately (i.e. theoretical coexistence). When co-occurring, competition can cause narrowing of the preferred range (i.e. exclusion), while mutualism can support range broadening (i.e. facilitation). When no interactions occur (e.g. due to completely different preferences with respect to other variables), no range changes are observed.*

Box 2.3: Terminology related to biotic interactions

All species interact with each other along the positive-negative continuum, leading to the introduction of specific terminology for each type of interaction. Basically, each interaction between two species can be classified as having a (1) positive, (2) negative or (3) neutral effect on the survival and reproduction of each individual species (Bronstein, 1994; Dodds and Whiles, 2010), hence resulting in a total of nine possible combinations. These combinations can be reduced to six unique interactions due to the inherent symmetry.

When both species are positively affected by the presence of the other species, the interaction is classified as **mutualism**. In contrast, when both species are negatively affected, the interaction is classified as **competition**. The combination where one species experiences a benefit and the other species experiences a detriment, the interaction is referred to as **parasitism** (or, alternatively, **predation** or **exploitation**). Some interactions do not provide any benefit or detriment for either species (i.e. neutral for both species) and are therefore classified as **neutralism**. When only one species benefits or suffers due to the interaction (without affecting the second species in any way), the interaction is classified as **commensalism** or **amensalism**, respectively.

Given the importance of plant interactions within terrestrial systems and the limited research performed on aquatic macrophytes (Brooker *et al.*, 2008; Callaway and Walker, 1997), more information is expected to be reported in future studies. This is imperative, as more experiments on these interactions (including field observations, replacement tests, laboratory experiments and phylogenetic research) are required to develop a biotic interaction filter (Guisan and Rahbek, 2011; Keddy, 1999; Pulliam, 2000). Additionally, considering the temporal dynamics of population and species characteristics, attention should be given to the potential effect of time and time-related variables, including season, life stage, size and density (Callaway and Walker, 1997).

It remains clear that, considering the relatively high number of potential interactions, natural observations provide more 'true' information than microcosm studies and tend to constitute a more accurate representation of the realised niche (Guisan and Thuiller, 2005). Nevertheless, this representation remains highly time-dependent and merely entails a snapshot of all ecological processes and interactions taking place within the considered timeframe (Araújo and Guisan, 2006; Lehmann, 1998). Hence, when aiming to estimate the intensity and direction of future distributions and interactions, experiments do provide the only alternative to expand currently existing trait matrices and to confirm (or reject) the ecological theory (Keddy, 2010).

2.3.2.3 Alien species

Besides the abovementioned challenges dealing with species-specific traits, natural dispersal and biotic interactions of the known native species pool, specific attention should be given to invasive alien species (IAS) (Hofstra *et al.*, 2020). Their presence is a direct result of increased globalisation and the ability to colonise unoccupied niches due to the possession of functional traits that differ in their value from native species (Perrings *et al.*, 2002; Thomaz and Cunha, 2010; Van Echelpoel *et al.*, 2016). More specifically, a discrepant dispersion method, resource uptake efficiency, rate of biomass production and the excretion of metabolic by-products allow IAS to outcompete and expel native species, thereby expanding the occupied niche (Zedler, 2003).

With current habitats changing at unprecedented rates, new niches are created continuously, allowing the establishment of and colonisation by IAS. Counteracting the impacts of IAS can be performed at pre-introduction (i.e. identification of invasive potential) or post-establishment (i.e. removal of IAS from colonised area) level (Early *et al.*, 2016), yet the invasive potential of many alien species still remains unknown and impedes the development of a priority list. So far, border control is by far the most implemented proactive management strategy to avoid the introduction of alien species and relies on several nationally and internationally renowned invasive species (Early *et al.*, 2016; IUCN, 2019). Yet, the inclusion of any alien species on these lists is often a mere reaction on reported detrimental effects elsewhere.

Observation-based HSMs allow to predict suitable habitats for IAS, though their reliability is questioned as (i) observations within new environments are not yet in equilibrium and (ii) observations within their native environment inherently include biotic interactions potentially absent within the new environment (Gallien *et al.*, 2012; Guisan and Thuiller, 2005). Hence, controlled experiments are required for determining the invasive potential, for instance via the functional response as applied by Dick *et al.* (2013), describing the increased resource-use efficiency of the invasive shrimp *Hemimysis anomala* compared to native mysid shrimps along a range of resource concentrations. In contrast, the assessment of competitive potential among macrophytes based on input-related comparisons remains limited, while output-based testing via the relative growth rate (RGR) is more common, e.g. Fagúndez and Lema (2019), Njambuya *et al.* (2011), Paolacci *et al.* (2018). Therefore, further research into the applicability of input-based approaches to determine the invasive potential of alien macrophytes is recommended, including comparisons among species that are phylogenetically close.

2.4 Contribution to the study objectives

The aim of this chapter was to create an overview of how specific biotic groups interact in shallow eutrophic freshwater systems and, from that, derive which biotic group(s) can provide a biological basis for developing a complex community. Throughout the chapter, attention was given to (1) the pollutant removal capacity and (2) the intra- and interspecies interactions of different biotic groups, both within the context of integrated constructed wetlands. Consequently, several key issues were identified to merit additional study to improve the implementation of ICWs, tackling societal, chemical and biological aspects. However, the main objective of this work deals with the biotic aspect of ICWs (see Section 1.2.1), which excludes both the societal and chemical aspects from further scrutiny.

The baseline for biotic development was provided in Section 2.3 and identified the contribution of macrophytes as a steering factor towards (1) increasing habitat complexity and (2) altering physicochemical conditions (see Section 2.3.2). By exerting these processes, macrophytes indirectly affect other biotic groups, including phytoplankton (e.g. shading, nutrient competition), periphyton (e.g. as substrate) zooplankton (e.g. as refuge area), macroinvertebrates (e.g. as food source) and fish (e.g. as refuge area). However, the presence of macrophytes is determined by matching environmental conditions and species-specific abiotic preferences. In order to derive these preferences and the habitats that comply to them, information from field data, laboratory experiments and expert knowledge is required. With this information, site identification and niche delineation can be automated by developing habitat suitability and species distribution models.

The development of HSMs and SDMs is a challenging task, requiring information on autecological processes, dispersal rates and biotic interactions. Particularly, attention is requested for the inclusion of their temporal dynamics, as prevailing conditions are continuously changing. Climate change, anthropogenic activities and the increased introduction of invasive alien species alter the environment both at small and large scale, hence resulting in changing communities, shifting niches and the potential extinction of specialist species throughout consequent years (Guisan and Rahbek, 2011; Pulliam, 2000; Vitousek *et al.*, 1997; Vos *et al.*, 2008).

As a result, pressure on ecological research increases as appropriate decision management requires the support of HSMs and SDMs to reliably forecast community changes caused by such environmental disturbance. Therefore, the remaining chapters will focus on the preferred abiotic conditions of macrophytes within wetland-like environments. Attention is given to species-specific preferences as well as management options to deal with non-native species (if necessary).

2.5 Conclusion

Construction of artificial treatment wetlands provides the opportunity to counter the ongoing loss of wetlands and related ecosystem services. Pollutant removal and presences of biotic components (phytoplankton, periphyton, zooplankton, macroinvertebrates, macrophytes and fish) have been reported within these systems, while an analysis of the biotic interactions highlighted the positive effect of macrophyte presence on ecosystem functioning. Yet, implementation is still impeded as specific integrated knowledge at the chemical and biological level is lacking. Therefore, a range of suggestions can be formulated to fill these knowledge gaps, being categorised in three domains: (i) societal, (ii) modelling and (iii) experiments. More specifically, within the societal domain more attention should be given to the inclusion of socio-economic expectations and needs when designing restoration projects. Secondly, developing abiotic habitat suitability models is called for to match environmental conditions with species-specific habitat preferences. Lastly, and most extensively, experiments are requested to improve understanding on (i) the functioning of constructed wetlands at the abiotic level (including the effects from external pressures and the impact on receiving water systems), (ii) species-specific temporal dynamics (including population processes and dispersal rates) and (iii) the applicability and effectiveness of pro- and reactive management when dealing with invasive alien macrophytes (including input-based indices and management effects).

3

Data-driven modelling for environmental data science[2]

Highlights

- Ecosystem management benefits from data-driven modelling
- No single-best method exists, but improvements are being made
- Decision trees are an accessible technique with acceptable performance
- Cross-validation helps to reduce overfitting and to increase data use efficiency

[2] This chapter is redrafted from Van Echelpoel, W.; Boets, P.; Landuyt, D.; Gobeyn, S.; Everaert, G.; Bennetsen, E.; Mouton, A. and Goethals, P. L. M. Species distribution models for sustainable ecosystem management in Developments in Environmental Modelling Vol. 27 (eds Y.-S. Park, S. Lek, C. Baehr and S. E. Jørgensen) Ch. 6, 115-134 (Elsevier, 2015).

Abstract

Ecosystems are characterised by complex interactions and a high degree of uncertainty due to their inherent dynamic behaviour. Model simulations help in decreasing these uncertainties and simultaneously create additional insight into existing ecological interactions. More specifically, species distribution models combine abiotic and species-specific information to describe current and simulate future species occurrence. These models derive their construction from data, knowledge or a combination of both, with the former being increasingly applied in ecological research related to conservation management and the effects of climate change. Here, five data-driven modelling techniques are discussed and compared to provide an overview of their strengths and weaknesses: decision trees, generalised linear models, artificial neural networks, fuzzy logic and Bayesian belief networks. From this overview, it becomes clear that no modelling technique is without drawbacks, making model selection often user- and case-dependent. Following model selection, data collection and preparation is highly technique-specific, including response balancing for decision trees and variable scaling for artificial neural networks. Moreover, model evaluation depends on the characteristics of the provided model output, providing most information when based on non-transformed observed or predicted response values. A shared challenge among the selected techniques consists of model regularisation by overcoming overfitting, which is partially tackled by implementing cross-validation or alternative approaches to improve data use efficiency. Overall, decision trees are relatively simple non-parametric techniques that allow for the integration of variable interactions, with random forests reporting promising results. The area under the receiver operating characteristic curve (AUC) represents a single-value and threshold-independent metric to assess model performance, while sensitivity (Sn) and specificity (Sp) provide potential as additional assessment metrics.

3.1 Setting the scene

In Chapter 2, the value of model development to estimate habitat suitability or species distribution has been highlighted in the context of ecological conservation. The majority of restoration projects counter the ongoing loss of biodiversity, yet suffer from high investment costs, short-term thinking, uncertain outcomes and insufficient inclusion of socio-economic needs and expectations (Catalano *et al.*, 2019; Diekmann and Featherman, 1998; Friberg *et al.*, 2017). Climate change adds to these uncertainties and challenges due to shifts in geographical range, seasonal activities, migration patterns and species interactions, while simultaneously increasing the risk of extinction for a large fraction of species (Braunisch *et al.*, 2013; IPCC, 2014; Walther, 2010).

Model simulations provide the opportunity to decrease some of these uncertainties and simultaneously create insight into existing ecological interactions. In this regard, the ability of models to extrapolate species distributions in space and time is a crucial contribution to maintaining and improving ecosystem structure and functioning. More specifically, these species distribution models (SDMs) allow to test biogeographic hypotheses (Leathwick, 1998), to fill in the gaps in current ecological knowledge (Ambelu *et al.*, 2014), to identify conservation areas and to determine invasion vulnerability (Domisch *et al.*, 2013; Hatten *et al.*, 2014; Sauer *et al.*, 2011).

SDMs are positioned along an axis between data-driven (empirical) and knowledge-based (conceptual) models (Dormann *et al.*, 2012; Mount *et al.*, 2016), though a single-best approach has not been identified due to the inability to create a universal grading of all existing models (Kampichler et al., 2010; Lawson et al., 2014). So far, data-driven models have been applied frequently when forecasting habitat suitability and species distributions (Elith and Graham, 2009; Marmion et al., 2009; Stohlgren et al., 2010).

Within this chapter, specific attention is given to a selection of five data-driven modelling techniques, being decision trees (DTs), generalised linear models (GLMs), artificial neural networks (ANNs), fuzzy logic (FL) and Bayesian belief networks (BBNs). Throughout the chapter, models are referred to as being species distribution models, as no strict assumptions on the available data are being made. However, the majority of data-driven models has been developed without the inclusion of dispersal dynamics or biotic interactions and is, therefore, defined as habitat suitability models (HSMs). Despite providing a valid alternative, knowledge-based models are built on known processes and are, therefore, considered to be out of the scope of this chapter.

The aim is to create an overview of frequently-applied modelling techniques and, in addition, to describe how to asses model performance prior to making predictions. By tackling these two objectives, an answer is provided to RQ1.3, defined in Chapter 1. Ultimately, this chapter concludes with a promising modelling approach for sustainable ecosystem management.

3.2 Model development procedure

In general, the model development procedure entails a sequence of successive steps to be performed. The number and focus of these steps differ among authors, which calls for a comprehensive standardisation, though allows the identification of several recurring steps. For instance, a list of ten successive steps is provided by Jakeman *et al.* (2006), while Guisan and Zimmerman (2000) only mention five important steps. Still, sequential steps are not always clearly separable and some can be combined in one overarching step (Austin, 2002; Jakeman *et al.*, 2006). Here, prior to applying the model for inferring predictions, four main steps are identified based on Guisan and Zimmerman (2000) and mentioned in Table 3.1: (1) create a conceptual framework, (2) collect and explore the data, (3) apply the most appropriate modelling technique and (4) calibrate the selected model and validate the model with independent data.

As prior knowledge is often limited and the initial goals of long-term studies and restoration projects often change (Catalano *et al.*, 2019; Friberg *et al.*, 2017), it is clear that careful design (i.e. "create conceptual model" in Table 3.1) and data collection (Step 2 in Table 3.1) are major challenges, for which a balance between robustness, general relevance, and specific needs has to be sought. Therefore, a careful, well-balanced combination of data, expert knowledge, and user convenience is recommended (Goethals, 2005), especially when developing process-based models.

Yet, both model design and data collection have become less significant during the past decades, as the unprecedented progress in data collection, storage and availability has supported a rise in the applicability and importance of data-driven models for decision-making (Benito *et al.*, 2013; Gibert *et al.*, 2018a). Still, the creation of a conceptual framework remains a valid step, though relatively more attention is (and should be) spent on data exploration and proper pre-processing (Zhang *et al.*, 2003).

Following model conceptualisation and data characterisation, model selection can be based on a series of objective parameters (e.g. performance measures in Table 3.1, Step 4), while additionally depending on the preference of the modeller (i.e. introduction of subjectivity) because no model can be considered as the best option in every situation (Gibert *et al.*, 2018b; Mount *et al.*, 2016; Mouton *et al.*, 2010). Consequently, several authors tend to combine multiple modelling techniques (i.e. "ensemble modelling") in order to predict future species distributions more reliably (Benito *et al.*, 2013; Domisch *et al.*, 2013; Gallardo and Aldridge, 2013; Thuiller, 2003).

Table 3.1: Summary of the four main steps in the ideal modelling procedure, including relevant literature.

Step	Goal	Relevant literature
1. Create conceptual model	Becoming aware of the situation to be investigated, i.e. suggesting a hypothesis, identifying the required data and selecting the most appropriate model.	Jakeman *et al.* (2006), Austin (2002), Guisan and Zimmerman (2000)
2. Data collection and exploration	Collecting the required data according to Step 1, followed by exploring the data and elimination of data that can inhibit proper model development.	Zuur *et al.* (2010), Guisan and Zimmerman (2000)
3. Model application	Applying the selected modelling technique (see Step 1).	Guisan and Zimmerman (2000), Leohle (1983)
4. Model calibration and validation	Estimating and fine-tuning of model parameter values to fit the provided data, including calculation of performance measures (i.e. model fit to independent data set).	Allouche *et al.* (2006), Fawcett (2006), Manel *et al.* (2001), Guisan and Zimmerman (2000), Fielding and Bell (1997)

3.2.1 Create conceptual framework: model selection

When relying on models for making predictions, one should be aware that models are a mere conceptualisation of the ecosystem under study and that, consequently, the obtained results carry a certain degree of uncertainty (Wilson *et al.*, 2011). Throughout this section, an assortment of empirical (data-driven) models is described in more detail. Selection of the models is based on reported applications in ecological literature and the work of Franklin (2010), who provides an elaborate description of decision trees (DTs) and generalised linear models (GLMs), as well as a concise introduction to artificial neural networks (ANNs) and generalised additive models (GAMs). Furthermore, Franklin (2010) describes fuzzy logic (FL) as an approach that holds a lot of promise to improve the usefulness of the habitat suitability index (HSI). Additionally, Bayesian Belief Networks (BBNs) are described as they are mentioned in the overview of Goethals (2005), listing decision trees, ANNs, fuzzy logic and BBNs as soft computing methods worth mentioning when dealing with modelling species distributions. Each of the following subsections describes one of these techniques (DTs, GLMs, ANNs, FL and BBNs) in more detail, refers to a more elaborate or mathematical description in literature and provides two examples in which the technique has been applied.

3.2.1.1 Decision trees (DTs)

Decision trees are hierarchical structures represented by a sequence of knowledge rules (Everaert *et al.*, 2011). Their construction is based on an iterative process of identifying the most informative predictor and the accompanying threshold value(s), thereby limiting the necessity to specify a relationship between explanatory and response variables on beforehand (De'ath and Fabricius, 2000; Fox *et al.*, 2017; Svitok *et al.*, 2016). The data set is split according to this threshold and the next iteration starts until a specific stopping criterion is satisfied. Ultimately, the final model is characterised by a specific number of nodes (i.e. knowledge rules) and leaves (i.e. branch ends), reflecting model complexity and allowing for a graphical representation. A distinction can be made between classification (categorical response) and regression (continuous response) trees. For instance, a hypothetical classification tree with two nodes and three leaves is depicted in Figure 3.1.

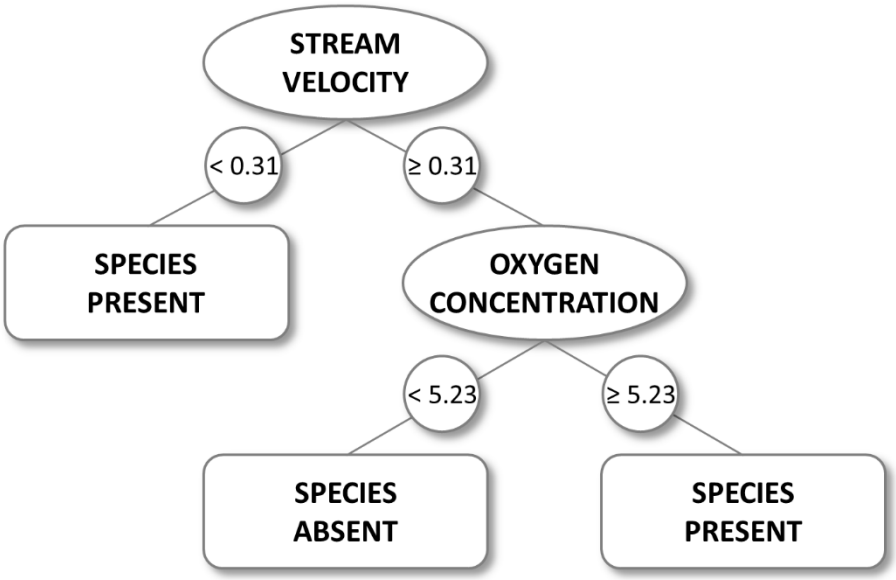

Figure 3.1: Illustration of a classification tree. *Species occurrence is determined by stream velocity and oxygen concentration and indicates the hierarchical importance of both predictors. The depicted model classifies a hypothetical species by using two nodes and three leaves.*

Decision trees have been frequently applied to model habitat suitability or species distribution, see for instance Boets *et al.* (2010), Boets *et al.* (2013b), Everaert *et al.* (2011), Hoang *et al.* (2010) and Van Echelpoel and Goethals (2018). Main advantages of decision trees are the comprehensibility of the model structure (e.g. Figure 3.1), since it closely resembles human reasoning (Kotsiantis, 2011), the ability to deal with relatively small datasets (Everaert *et al.*, 2011) and the possibility to identify (non-linear) interactions between predictors (Franklin, 2010; Svitok *et al.*, 2016). More information on decision trees can be found in Rokach (2008).

Examples

Decision trees have been successfully applied to determine the presence of alien macrocrustaceans in surface waters in Flanders (Boets *et al.*, 2013b). Both classification and regression trees were developed in order to describe species distribution (present/absent) and both richness and abundance (continuous response variables), respectively. In short, they concluded that presence and species richness of macrocrustaceans are likely to increase with improving water quality, probably accompanied by a slight decrease in abundance of the most dominant alien taxa (Boets *et al.*, 2013b). Useful applications of the inferred knowledge on these alien species include management planning and investment decisions, which are highlighted by the United States National Management Plan on invasive species (Kolar and Lodge, 2002).

In vegetation ecology, regression trees have been applied to describe the potential migration of trees under changing climatic conditions (Iverson and Prasad, 1998). Among the selected species, Iverson and Prasad (1998) observed different responses to climate change with an additional remark that future redistributions will be dependent on migration rates through fragmented landscapes. This application fits in the idea that climate change will eventually lead to a large redistribution of tree species considering the increase in average surface temperature and the change in precipitation patterns (IPCC, 2014; Kundzewicz *et al.*, 2014).

Additional remarks

Despite their comprehensibility, classification trees are not always the best option in terms of model performance. In comparison with other modelling techniques, decision trees have shown to perform better (Boets *et al.*, 2013a) and worse (Hoang *et al.*, 2010), illustrating the case-dependency of model performance. General drawbacks of decision trees are related to their instability (an error in the top split will propagate down to all splits below (Franklin, 2010; Hastie *et al.*, 2009)), the limited incorporation of external ecological knowledge and the possibility of overfitting the model.

These drawbacks tend to limit the applicability of basic decision trees on external or independent data sets, yet the development of more advanced tree-based models (e.g. boosted regression trees, random forests (see Box 3.1)) has countered most of this criticism by reporting the outperformance of other modelling techniques (Breiman, 2001; Marmion *et al.*, 2009; Stohlgren *et al.*, 2010). Furthermore, when dealing with high amounts of data, large grown trees can be obtained, which are, due to their complexity, difficult to interpret. Pruning, which is the removal of one or more sub-trees to avoid overfitting, weights model complexity versus proximity to the data (model fit). By allowing (small) errors, trees will be less complex and the obtained rules are considered more generally applicable and improve the regularisation of the developed model (Mingers, 1989).

Box 3.1: Random forests as an ensemble of decision trees

The high sensitivity of decision trees towards erroneous data and the possibility towards overfitting the data have resulted in a variety of alternative decision tree configurations. Among these, Breiman (2001) introduced the possibility to combine several individual trees into a single model (i.e. an ensemble model), which averages the overall model response and thereby limits the model's sensitivity towards errors and overfitting.

To avoid a strong correlation of the individual trees, instances are randomly selected from the provided training data for each tree. Subsequently, within each tree a random sub-selection of the available variables is made (i.e. the square root of the number of variables, by default) prior to defining the node-specific threshold value. Due to this approach, a fraction of the training data remains unused for each tree, which is applied to infer a tree-specific *out-of-bag* performance estimate. These estimates can be pooled to provide an overall evaluation of model performance. Alternatively, a completely independent data set can be used to perform external model validation. For each instance within this data set, the response of all individual trees is averaged and can be reported as a fractional distribution or a single response (if a specific threshold value is provided). The development of a random forest is visually represented in Figure 3.2.

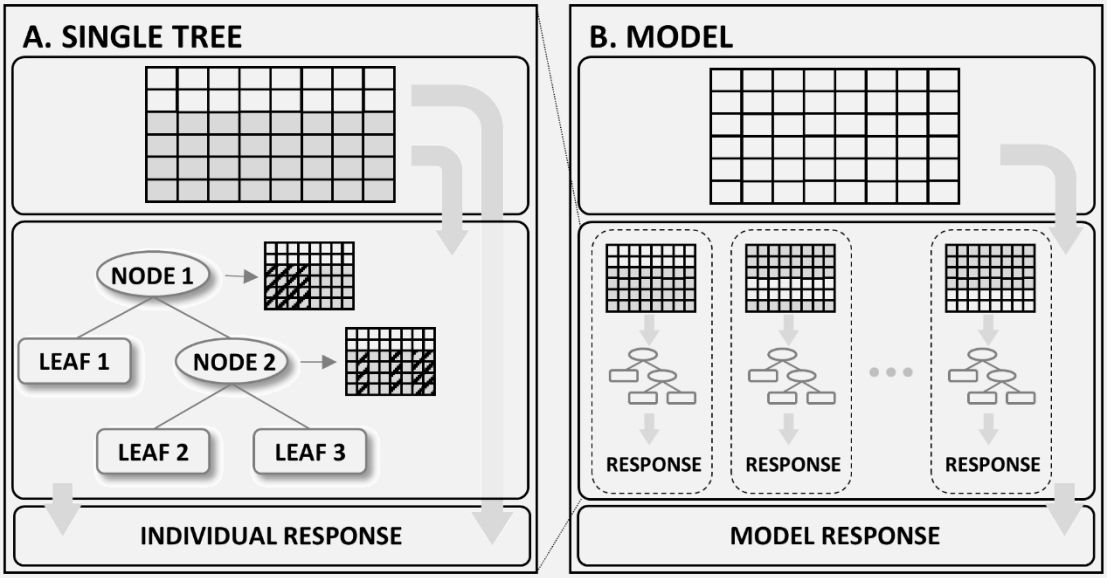

Figure 3.2: Development of a random forest. *The final model consists of a predefined number of individual trees, which are all unique due to the variation in the provided training data. A: Development of a single tree with a fixed data set and varying variable selection for each node; B: Development of the model with the original data and the random instance selection per individual tree.*

3.2.1.2 Generalised linear models (GLM)

GLMs are a generalisation of ordinary linear regression models and are based on three elements: (1) a random component that assumes a probability distribution of the response variable *Y* (e.g. exponential, binomial), (2) a systematic component specifying the predictors in a linear form with their respective coefficient and (3) a link function describing the relationship between the former two elements (*random component = link function*(*systematic component*)) (Zuur *et al.*, 2009). The predictors used for the systematic component can be independent predictors of higher order (e.g. *velocity*^{2}) to model curvilinear effects or an interaction of predictors (e.g. *depth·oxygen*) (Willems, 2010; Zuur *et al.*, 2009). The mathematical expression for GLMs is conceptualised in Equation 3.1 for a single response variable (*Y*).

$$g^{-1}[E(Y|\boldsymbol{X})] = \beta_0 + \sum_{j=1}^{k} \left(\beta_j X_j\right) + \varepsilon \qquad \text{(Equation 3.1)}$$

With g^{-1} the inverse link function, Y the response variable, $E(Y|\boldsymbol{X})$ the expected distribution of Y conditional to the set of predictors ($\boldsymbol{X} = [X_1, X_2, \dots, X_k]^T$), X_j the jth predictor (out of k predictors, including higher order and interaction terms), β_0 the intercept, β_j the slope related to the predictor X_j and ε the remaining error.

GLMs are regularly used in ecology to predict and describe the behaviour of a continuous response variable (e.g. abundance, probability of occurrence) in relation to environmental predictors, see for instance Ambelu *et al.* (2014), Everaert *et al.* (2014), Guisan *et al.* (2006) and Thuiller (2003). Important advantages that are related to GLMs include the ability to handle different types of distribution for the response variable, the possibility of constraining the predicted response variable in a certain range (e.g. between 0 and 100 % probability of occurrence) with statistical substantiation and the incorporation of potential solutions (by using extensions) to deal with overdispersion (i.e. variance of the data is larger than the intrinsic variance of the anticipated distribution (Davison, 2001)) (Guisan *et al.*, 2002).

GLMs are, as mentioned above, limited to the assumption that the response variable is linked to a linear combination of all predictors (see Equation 3.1) (Guisan *et al.*, 2002; Zuur *et al.*, 2009). An extension of GLMs assumes that when the predictors are smoothed by using a smoothing function, the linear combination of these functions is linked to the response variable. This extension is referred to as generalised additive models (GAMs) and is able to deal with non-linear, non-monotonic relationships between the predictors and response variables (Guisan *et al.*, 2002). The mathematical expression of GAMs is conceptualised in Equation 3.2 for only one response variable (*Y*). More information on GLMs and related extensions (e.g. generalised additive models (GAMs), generalised linear mixed models (GLMMs), generalised additive mixed models (GAMMs)) can be found in Zuur *et al.* (2009).

$$g^{-1}[E(Y|\boldsymbol{X})] = \beta_0 + \sum_{j=1}^{k} f_j(X_j) + \varepsilon \qquad \text{(Equation 3.2)}$$

With g^{-1} the inverse link function, Y the response variable, $E(Y|\boldsymbol{X})$ the expected distribution of Y conditional to the set of predictors ($\boldsymbol{X} = [X_1, X_2, \ldots, X_k]^T$), X_j the jth predictor (out of k predictors), β_0 the intercept, f_j the smoothed function related to the predictor X_j and ε the remaining error.

Examples

The abiotic preferences of aquatic macroinvertebrates in tropical river basins was assessed by Everaert *et al.* (2014), who used logistic regression models (LRM), being a specific type of GLMs. In this study, LRMs were used to deduct relationships between abiotic variables and species presence in three tropical river basins (Ecuador, Ethiopia and Vietnam). Constraining the response variable between 0 and 1 (i.e. 0 and 100 % probability of occurrence) allows future application of the developed model outside the observed predictor range (e.g. future environmental conditions), while still resulting in a plausible response variable.

In vegetation ecology, GAMs were developed in order to describe and predict the distribution of the Aleppo pine (*Pinus halepensis*) in Europe (Thuiller, 2003). Considering GAMs to apply a smoothing approach, no interaction terms have to be included, which provides an advantage over GLMs. The results showed a northward expansion of *Pinus halepensis* with minor contractions in southern Europe as a consequence of future climate change (Thuiller, 2003). As already mentioned, dispersion of trees due to changing climate conditions will also be affected by the possibility and rate of migration through fragmented landscapes (Iverson and Prasad, 1998), which can limit their dispersal and eventually influence the overall carbon cycle.

Additional remarks

GLMs and classification trees were both applied to predict the presence of four vegetation alliances in the Mojave Desert (California). The application of GLMs to classify the considered vegetation alliances as present or absent resulted in a lower classification accuracy with the training data, but performed relatively better on an independent data set (Miller and Franklin, 2002). Similarly, GLMs and GAMs performed worse compared to random forests (a specific type of decision trees) when being applied to predict the effect of climate change on both native and invasive species (Gallardo and Aldridge, 2013). Drawbacks of GLMs are related to the assumption of the response variable being linked with a linear combination of the predictors, the possibility of overdispersion with binomial- and Poisson-like data (Venables and Dichmont, 2004) and the assumption that the response variable is characterised by a specific distribution. Several of these issues are tackled with GAMs and GLMMs, though these are simultaneously characterised by an increased mathematical complexity.

3.2.1.3 Artificial neural networks (ANN)

Artificial Neural Networks (ANNs) are non-linear mapping structures that resemble the human brain (Lek and Guégan, 1999) or, more specifically, the neurons present in it (Basheer and Hajmeer, 2000). A combination of predictors is handled by a sequence of neurons and will ultimately lead to the response variable (see Figure 3.3). As a consequence, ANNs are considered to be a 'black-box' approach (Lek and Guégan, 1999) that use predictors to infer the state of the response variable without reporting intermediate predictor combinations and transformations. ANN application in ecology remains limited, though includes some success stories, see for instance Brosse *et al.* (1999), Dedecker *et al.* (2004), Goethals *et al.* (2007) and Thuiller (2003). Important advantages are related to the high tolerance for noise and measurement errors and the ability to recognise relations between predictors and response variables without ecological knowledge and regardless of the system's non-linearity and the problem's dimensionality (Basheer and Hajmeer, 2000). More information related to ANNs can be found in Zurada (1992), while practical applications in supporting river management are available in Goethals et al. (2007).

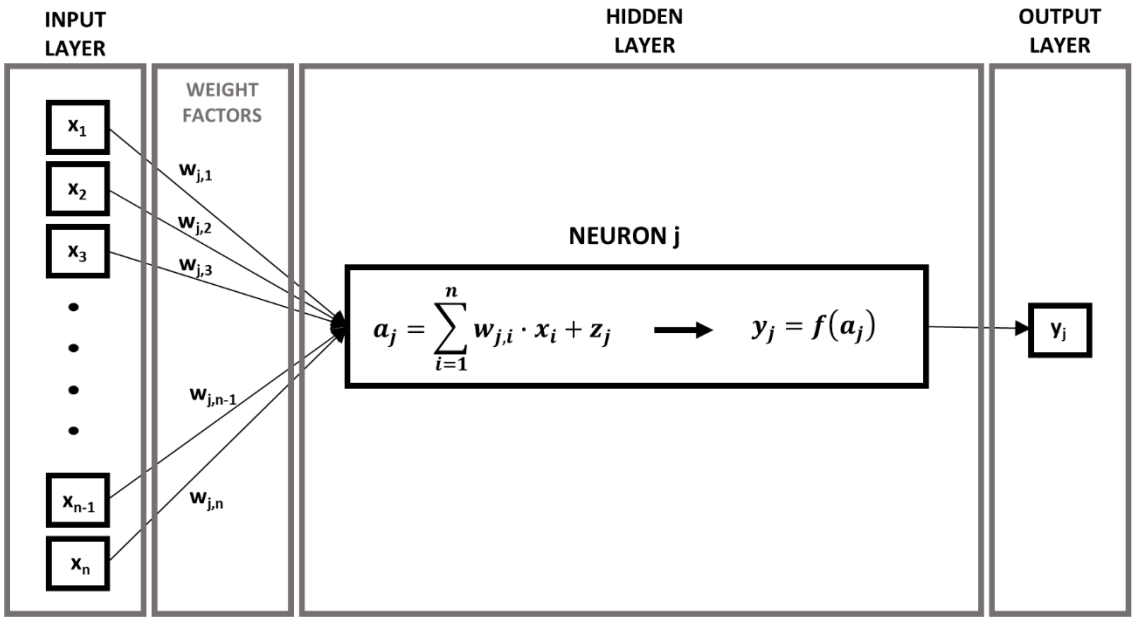

Figure 3.3: Schematic illustration of a single neuron in a single hidden layer ANN. *Input values are received from* n *predictors (*x*), associated with a specific weight (*w_j*) and an overall bias term (*z_j*). A new variable (*a_j*) is calculated and transformed by a transfer function f, resulting in the* j*-th output (*y_j*).*

In ecology the most popular types of ANNs are Kohonen self-organizing maps (SOM) and backpropagation networks (BPN), among which the latter are frequently used (Goethals, 2005; Lek and Guégan, 1999). BPNs are multi-layer feed-forward neural networks (also called 'multi-layer perceptron', MLP) in which the information flows unidirectionally. The network connects the predictors with the response variables through a number of hidden layers, which are successively arranged and contain the neurons, being non-linear elements. The neurons present in the hidden layers create new 'variables' based on the predictors or variables from a previous layer, multiplied with a variable-specific weight factor and the addition of a bias term (see Figure 3.3 in case of a single hidden layer with a single neuron). In a BPN there are no lateral connections (i.e. between neurons of the same layer), nor feedback mechanisms.

Examples

Olden *et al.* (2006) acknowledged the presence of complex interactions in aquatic communities and applied ANNs to approach the existing hierarchic structure. By considering the presence of different spatial scales (i.e. valley-scale, watershed-scale and river-scale) and the related creation of nested ANNs, the ability to introduce a limited amount of knowledge is illustrated. Based on this approach, Olden *et al.* (2006) observed that among the selected environmental predictors, climate variables have the highest mean importance. Consequently, when considering climate change in the near future, a change in the composition of currently existing communities can be expected.

Similarly, ANNs were applied by Dedecker *et al.* (2004) to describe and predict the habitat suitability of macroinvertebrate taxa in the Zwalm River (Belgium). They observed that different model structures result in different response variable curves describing the probability of presence in relation to dissolved oxygen. Furthermore, these macroinvertebrates are generally regarded as a proxy for overall water quality, and will, in light of climate change, be influenced by changing water quality due to altered hydrological systems (IPCC, 2014; Kundzewicz *et al.*, 2014).

Additional remarks

Brosse *et al.* (1999) compared the capacity of ANNs to fit observed patterns with multiple linear regression (MLR) and concluded that ANNs were more suitable due to the shortcomings of MLR related to higher levels of ecological complexity. A similar conclusion was reported by Brey *et al.* (1996) when comparing ANN and MLR for predicting production-to-biomass ratios. However, in another case, Willems (2010) observed that, when parsimony is considered important, GLMs were superior to ANNs. Drawbacks of ANNs are its behaviour as a black box model, a lack of fixed guidelines for optimal ANN architecture and limited inclusion of ecological concepts and relations (Basheer and Hajmeer, 2000; Brosse *et al.*, 1999; Thuiller, 2003).

3.2.1.4 Fuzzy logic (FL)

Fuzzy logic models are based on the assumption that a crisp classification of observations is not always straightforward and ecologically sound (Adriaenssens *et al.*, 2004a). When dealing with classification, one can use strict boundary conditions, e.g. when temperature is below 10 °C it is considered as 'cold', in between 10 and 20 °C as 'moderate' and above 20 °C as 'warm'. This results in a decrease of the number of response variables and a loss of information. Fuzzy logic allows the presence of an intermediate state in which the discretised variable (regardless of being a predictor or response variable) can belong to several classes with a certain membership (Mouton *et al.*, 2011). This overlap is described by a weight (membership) factor (between 0 and 1) of which the sum always equals 1 (see Figure 3.4). The resulting trapezoidal shapes depict the membership functions, whose shape can differ based on the type of response variable. A more detailed mathematical description can be found in Mouton (2008).

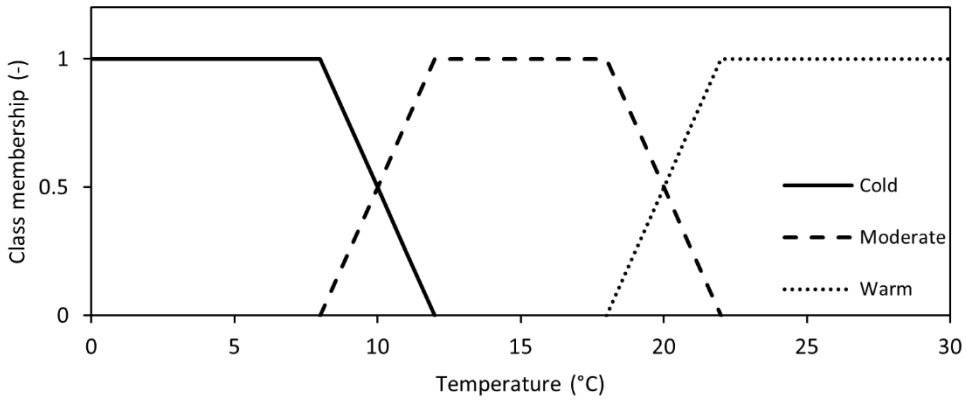

Figure 3.4: Concept of the fuzzy logic approach, illustrated with class membership in function of temperature. *The different classes (*Cold*,* Moderate *and* Warm*) are not crisp sets but are characterised by overlap between consequent classes. Class membership describes the weight of each class at a certain temperature and always sums to 1.*

Fuzzy logic is based on the construction of IF-THEN rules, extended with one or more AND-rules. For instance, IF temperature is high AND oxygen is high AND … THEN respiration is high. Each of these fuzzy rules generates an output and an accompanying fulfilment degree that takes into account all membership degrees of the predictors (e.g. minimum, maximum, product). Afterwards, these individual outputs and fulfilment degrees are combined to determine the global fuzzy output. For instance, Mamdani-Assilian models are linguistic fuzzy models that apply t-norms to determine the individual and global fulfilment degrees (Assilian, 1974; Mamdani, 1974), illustrated in Mouton (2008) and Van Broekhoven and De Baets (2008). A simplified version with two predictors and a minimum-based aggregation is depicted in Figure 3.5.

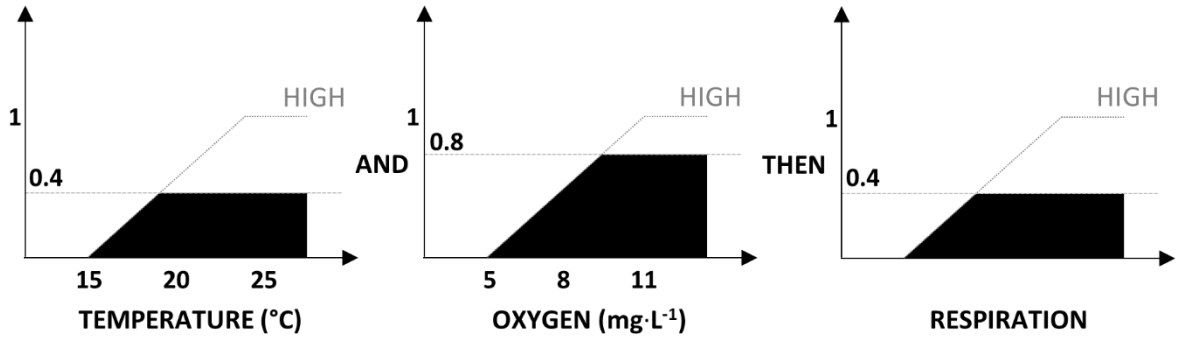

Figure 3.5: Membership determination in the response variable in fuzzy logic. *The response class is determined by temperature and oxygen, which are both characterised as* High *with a specific membership (i.e. 0.4 and 0.8, respectively). Calculation of the membership to the* High *class in respiration is here determined as the minimum membership of the two predictors. The represented IF … THEN … rule depicts a hypothetical classification of the respiration.*

Finally, the resulting membership degrees can be handled in two different ways: (i) defuzzification and (ii) by a fuzzy classifier. Defuzzification of Mamdani-Assilian models considers the global fuzzy output in combination with the accompanying fulfilment degrees and the subsequent conversion into a single response value (e.g. mean of maximum, center of gravity (Van Broekhoven and De Baets, 2006)). The second approach entails normalisation and converts the different membership degrees into values of which the sum equals one (Van Broekhoven *et al.*, 2006). The membership to each possible response variable class is described by this set of values.

After being developed in 1965 (Zadeh, 1965), the fuzzy set theory has been adopted by ecology, though remains scarcely applied, see for instance Adriaenssens *et al.* (2004a), Fukuda *et al.* (2011), Mouton *et al.* (2008) and Salski (1992). Important additional advantages include the potential decrease of complexity by combining a range of response variables in a single class and the possibility to include expert knowledge. The latter influences the classification of predictors, the shape of the membership functions and the rules, ultimately resulting in a more ecologically sound model.

However, expert knowledge is not an exclusive requirement for applying fuzzy logic, since both rules and fuzzy sets can be identified from data by means of fuzzy clustering, neural learning methods or genetic algorithms (Gobeyn *et al.*, 2017; Guillaume, 2001). This is specifically applied for numerical models (referred to as Takagi-Sugeno models) that focus on accuracy (Mouton, 2008). When models are based on predictors and response variables partitioned in classes, one speaks of linguistic fuzzy models. More information on fuzzy logic can be found in Klir and Yuan (1995).

Examples

Based on fuzzy logic a model was developed to predict the effects of different management options on a river and the accompanying influence on the spawning options of the European grayling (*Thymallus thymallus*) in the Swiss river Aare (Mouton *et al.*, 2008). This case illustrates the advantage of being able to combine expert knowledge with data in order to compensate for situations in which insufficient data is collected. Hence, data-driven techniques can help to mitigate bottlenecks related to knowledge-based rule-setting, which is considered to be time consuming and complex (Mouton *et al.*, 2008). Furthermore, this combination of data and expert knowledge allows to use predictor data with a specific uncertainty, as is the case when using simulated future environmental conditions as predictors.

Similarly, fuzzy logic was applied to evaluate habitat suitability of topmouth gudgeon (*Pseudorasbora parva*), an invasive fish species in Japan (Fukuda *et al.*, 2011). Several types of predictors (e.g. river width, canal network index, residential area, etc.) were implemented in the model structure, which illustrates the ability of fuzzy logic to deal with a variety of predictors. However, adding predictors also requires the definition of predictor-specific membership degrees and additional fuzzy rules. On the other hand, when future conditions result in predictor values outside the observed range (e.g. increased river width due to altered hydrological systems (IPCC, 2014)), predictions of distribution patterns can still be made due to the incorporation of expert knowledge in the original model.

Additional remarks

Fuzzy logic models have shown to perform similarly when compared with random forests (a specific type of decision tree), although when considering transparency, fuzzy logic models scored better because of their ability to combine ecological relevance with reasonable interpretability (Mouton *et al.*, 2011). Drawbacks of fuzzy logic are the increase in complexity with increasing amount of predictors (Ahmadi-Nedushan *et al.*, 2006), the loss of information due to data discretisation and the possibility that the implementation of expert knowledge rules is both cost- and time-intensive (Kompare *et al.*, 1994).

3.2.1.5 Bayesian belief networks (BBNs)

Bayesian Belief Networks (BBNs) are multivariate, probabilistic models that consist of a directed acyclic graph wherein nodes represent discrete variables and arrows causal relations (Aguilera *et al.*, 2011). Probability distributions quantify the probability of a variable being in one of its discrete states given the states of the preceding nodes in the graph (i.e. conditional probability). This way, uncertainties are explicitly accounted for and can be propagated from predictor to response variable using the rule of Bayes. Consequently, the output of a BBN is not a single value but a probability distribution over the states of the response variable.

BBNs have been applied in ecology to model species distributions, see for instance Keshtkar *et al.* (2013), Marcot *et al.* (2001), Pollino *et al.* (2007) and Smith *et al.* (2007). Important advantages of this modelling approach include the ability to update conditional probabilities when new knowledge is available (Castelletti and Soncini-Sessa, 2007), high model transparency, the potential to deal with missing data and the ability to complement empirical data with expert knowledge (Landuyt *et al.*, 2013). By modelling the joint probability distribution over all considered variables (both predictor and response variables), BBNs differ from most other modelling techniques that only focus on accurately predicting the response variable. More information on BBNs can be found in Jensen and Nielsen (2007).

BBNs can be developed purely data-driven by using data to infer both the network structure and the conditional probability tables (CPTs). However, generally, the structure of the network is based on expert knowledge, while the CPTs are based on data (Landuyt *et al.*, 2013). Although such partially knowledge-based models may accurately represent the ecological functioning of the system based on current knowledge, they are often outperformed by purely data-driven models. For optimal classification performance (e.g. presence/absence models), several simple graph structures, such as, naive bayes (NB) classifiers and tree-augmented naive bayes (TAN) classifiers, have been proposed (Aguilera *et al.*, 2010; Friedman *et al.*, 1997). The causal links in NB classifiers are limited to direct links from the response variable to each predictor variable, while TAN classifiers also allow causal links among predictor variables mutually. Although these models usually do not grasp all dependencies and independencies of the system being modelled, they generally perform well in classification tasks (Friedman *et al.*, 1997).

Examples

A BBN has been developed by Marcot *et al.* (2001) to determine the effect of different land management alternatives on the habitat and population viability of fish and wildlife that were at risk. They observed that BBNs can be easily applied for modelling the effect of planning alternatives on fish and wildlife and that they are an interesting decision support tool. In this case, the application of BBNs is considered as a complementary tool since sufficient empirical data is provided to determine the effect of different land management alternatives. In case sufficient empirical data is lacking (e.g. altered landscapes and future environmental conditions), BBNs allow to perform risk assessments based on the reported likelihoods.

Besides being applied for determining land management issues, BBNs can also be used to model the effects of different catchment management alternatives on limiting the current degradation of water quality (Keshtkar *et al.*, 2013). By including stakeholders and expert judgment, Keshtkar *et al.* (2013) optimised the preliminary model, constructed CPTs when qualitative data was lacking and validated the results. Their results showed that riparian restoration has an important influence on overall water quality even when considering the cost of implementation (Keshtkar *et al.*, 2013).

Additional remarks

BBNs are comparable to ANNs as both techniques rely on a network approach. However, compared to ANNs, BBN models are more transparent, enable the integration of expert knowledge and require less data (Landuyt *et al.*, 2013). Therefore, BBNs are more suitable for participatory model development and validation. Additionally, the model structure itself can be used as a decision support tool considering the visual representation of causal relationships in an environmental situation.

Two studies compared the predictive performance of BBNs with other modelling techniques and concluded that the predictive performance of BBNs is relatively good compared to ANNs and fuzzy logic models (Adriaenssens *et al.*, 2004b) and compared to logistic regression (Ordóñez Galán *et al.*, 2009). Drawbacks of BBNs include the difficulty to implement temporal dynamics and information loss through discretisation of continuous variables. Although advanced model types exist to deal with temporal dynamics (e.g. time-sliced models, see Kjærulff (1995)) and continuous variables (e.g. hybrid Bayesian networks, see Aguilera *et al.* (2010)), other modelling techniques may be more suitable.

3.2.1.6 Summary of advantages and drawbacks

A summary of the advantages and drawbacks of the selected modelling techniques is provided in Table 3.2. General drawbacks of each approach are mentioned despite the existence of several recently developed techniques that, at least partially, compensate for these weaknesses. However, most compensating techniques have a negative influence on the main advantages, which highlights the need of a well-balanced and carefully considered implementation.

Table 3.2: Summary of model advantages and drawbacks. *An overview is provided of the five modelling techniques (decision trees (DT), generalised linear models (GLM), artificial neural networks (ANN), fuzzy logic (FL) and Bayesian belief networks (BBN)) discussed in previous subsections.*

Technique	Advantages	Drawbacks
DT	- Transparent modelling technique; - Able to deal with small data sets; - Able to identify interactions between explanatory variables; - No need to define relationships or distribution in advance.	- Limited incorporation of knowledge; - Potentially vulnerable to overfitting; - A single tree can provide unstable results; - Large datasets can lead to large, complex trees.
GLM	- Easy to use; - Useful for specific problems, e.g. predicting probability of occurrence with statistical substantiation.	- Limited incorporation of knowledge; - Assumes the presence of specific distribution of the response variable.
ANN	- High tolerance for noise and measurement errors; - The ability to recognise relations between predictors and response variables when knowledge on the system's functioning is lacking.	- Acts as black box model; - Lack of guidelines for optimal design; - Low ecological relevance; - Limited explanatory power.
FL	- Absence of strict boundary values; - Ability to complement empirical data with expert knowledge; - Ability to incorporate uncertainty scenarios (e.g. climate change) by possibility approach.	- Increased complexity with increasing number of predictors; - Information loss due to data discretisation; - Construction of knowledge-based rules is time intensive.
BBN	- Accounts for uncertainties explicitly; - Ability to incorporate uncertainty scenarios (e.g. climate change) by probability approach; - Straightforward propagation of uncertainties related to model inputs; - Ability to complement empirical data with expert knowledge.	- Inability to implement temporal dynamics; - Information loss due to data discretisation; - Construction of knowledge-based rules is time intensive.

3.2.2 Data collection and exploration

Following modelling technique selection based on abovementioned advantages and drawbacks, data is to be collected for model training (Step 2, Table 3.1). Environmental observations collect information on a myriad of variables, often classified as explanatory and response variables. Typically, explanatory variables include all variables to be considered to explain the observed pattern within the response variable of interest and can be biotic and abiotic. However, the majority of HSMs focuses on defining suitable abiotic conditions and thereby restricts the extent of the explanatory variable space.

Variables can be discrete or continuous, with the former representing a limited number of possibilities (e.g. 5 different classes of land use), while the latter is not characterised by fixed thresholds to distinguish classes (e.g. river width expressed in meters). Most HSMs aim to accurately predict habitat suitability of a single species, thereby relying on a presence/absence statement (discrete) or a measure of abundance (continuous) in the response variable. Typically, models developed with a continuous response variable tend to be more sensitive compared to presence-absence models, despite containing potential biases related to seasonality, long term fluctuations and different sampling techniques (Ysebaert et al. 2002).

Still, the majority of HSMs is trained with a discrete response variable, ranging from the basic presence-only (PO) to completely presence-absence (PA). Presence-only data sets describe the locations where a specific species is observed, occasionally making use of records from museums or herbaria (Graham et al. 2004), though without providing any information on unsuitable conditions (Ward et al. 2009). In contrast, presence-absence data include information on species absences, yet these do not necessarily reflect effective absences. More specifically, reported absences combine true and false absences, the latter of which is composed of species being present without being observed (non-detectability) and species being absent due to historical or dispersal limitations (future potential) (Anderson and Raza, 2010). These false absences negatively affect model accuracy by providing ambiguous information (Lobo *et al.*, 2010).

Obtained data sets are rarely perfect and often contain one (or more) variables with missing values, erroneous notations, redundant variables and an unbalanced response variable (Gibert *et al.*, 2018a). Data exploration and pre-processing are therefore crucial tools to characterise and improve the quality of the obtained data and, thereby, increase the reliability of model outcomes (Zhang *et al.*, 2003; Zuur *et al.*, 2010). Data exploration allows to obtain a graphical representation of a variable's distribution, with boxplots, dotplots and histograms being frequently applied to identify potential deviating instances (Zuur *et al.*, 2010). Transformation or removal of these outliers are common approaches to improve data quality, yet the lack of uniform guidelines cause it to be relatively subjective and open to further study.

Similarly, predictor assessment by means of a correlation index or principal component analysis (PCA) helps in identifying explanatory variables that contain similar information. These are helpful in understanding ecological interactions and processes, yet potentially compromise model development due to limited information gain compared to the computational cost. Reduction of data dimensionality by correlated variable removal or the creation of a new set of independent variables based on the PCA axes generally supports the development of simpler and more transparent models (Guo *et al.*, 2015; Wilson *et al.*, 2011). The intensity of these effects depends on data characteristics and varies among modelling techniques.

Aside from abovementioned pre-processing, additional changes are potentially required prior to model training. For instance, ANN requires predictors to be rescaled to a predefined interval, ranging between 0 (or -1) up to 1, in order to make reliable predictions. Without this rescaling, predictors with an extensive range can have a higher influence, which can be artificially altered by changing the unit. Similarly, balancing of the response variable within the training data is highly recommended for DTs in order to avoid model preference towards the class with the highest frequency. To this end, a balanced ratio can be obtained via (i) random subsampling of the class(es) with higher abundance (Araújo and Guisan, 2006), (ii) oversampling of the class(es) with lower abundance or (iii) a combination of both (see Figure 3.6).

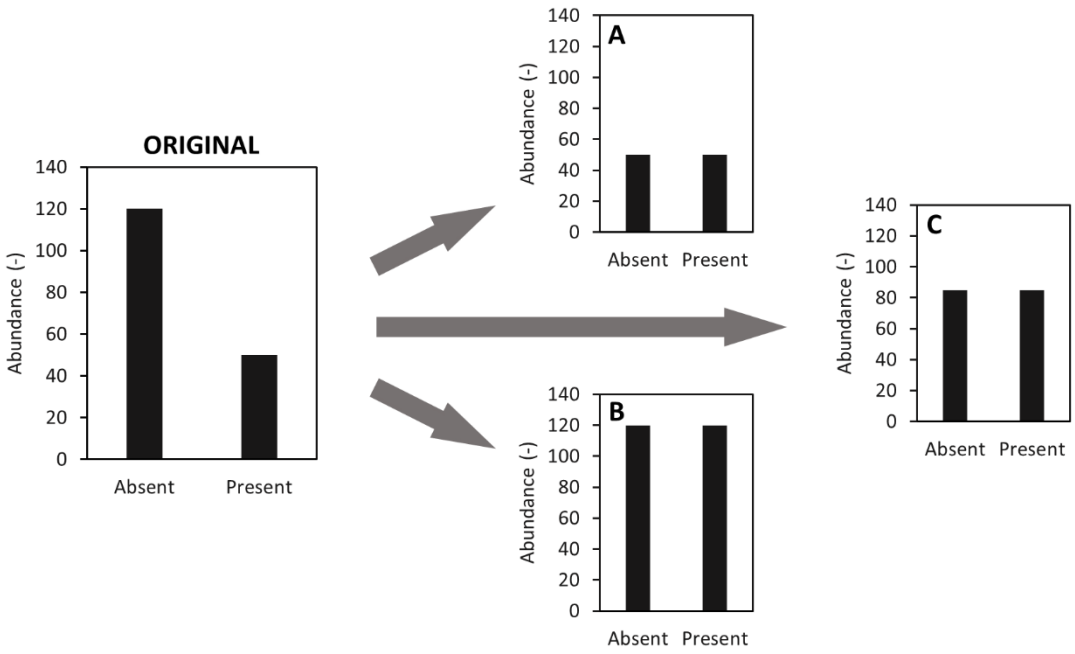

Figure 3.6: Balancing of data describing the occurrence of a non-specified organism. *Occurrence was assessed at 170 locations and resulting in 120 absence statements and 50 presence statements. A balanced dataset is created by A: randomly omitting data related to the absence of the organism (subsampling); B: randomly duplicating data related to the presence of the organism (oversampling) or C: applying a combination of subsampling and oversampling.*

3.2.3 Model application

A variety of parameters is linked with model development, distinguishing between algorithm parameters (or 'hyperparameters') and model parameters (e.g. regression coefficients). Hyperparameters are often subjectively defined prior to model training and remain unaltered regardless of the provided data, while model parameters are an intrinsic element of the final model and highly dependent on the training data. Subjective selection of hyperparameter settings can affect model performance drastically, hence preliminary optimisation is highly recommended. Ideally, all potential hyperparameter settings are tested to identify the best-performing combination(s), though this number tends to increase exponentially with every additional hyperparameter to be considered. Alternatively, random selection of a subset (e.g. 60 combinations) provides a first overview of potential performance and identifies a starting point for hyperparameter optimisation, while being generally faster than the traditional grid search (Bergstra and Bengio, 2012). This procedure reduces overall calculation time as it does not require for all combinations to be assessed, yet risks that the global optimal hyperparameter combination will not be found.

3.2.4 Model calibration and validation

The last step in the model development procedure entails the calibration and validation of the model (Table 3.1). During calibration, the training data is used to update model parameters in order to improve model fit, providing splitting values for DTs, coefficients of GLMs, weights in ANNs, inflection points in FL and CPTs in BBNs. Calibration is run until a specific stopping criterion is met (e.g. number of nodes in DTs, number of layers in ANNs, numerical error between observations and predictions). Defining this criterion is part of deciding hyperparameter values and tends to differ in function of the intended model use. For instance, descriptive models aim for a close model fit (thus a higher complexity), while predictive models are more general to allow transferability.

Following calibration, the model is validated by assessing the discrepancy between model predictions and observations, relying on internal or external validation. Internal validation compares the observations with the predictions made for the training data, though is considered to be insufficient for model validation as it does not allow to assess model performance objectively (Araújo *et al.*, 2005a). Therefore, an external data set is preferred to test the model's generality and report its performance more objectively (Dormann *et al.*, 2012). However, completely independent data (e.g. data that differs at spatial and/or temporal level) is rarely available and is often replaced by pseudo-independent data by means of randomly subsampling the original data set into a training and validation set (i.e. the 'holdout' method, see Figure 3.7) (Araújo *et al.*, 2005a). Based on this final comparison, model performance can be estimated. For the remainder of this section, attention is given to the different validation metrics and techniques.

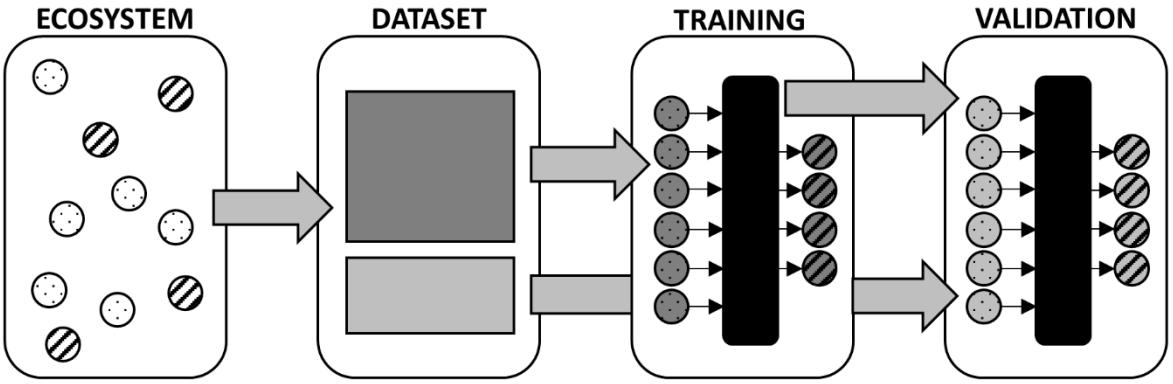

Figure 3.7: Development of a data-based model. *The first step (left) describes the observation of the ecosystem, resulting in a measurement step. The holdout method requires a part of the data set to be separated (the validation set), while the other part is used for model training (third step). After model training, validation is performed (right) by comparing the predicted and actual response values of the validation set.*

3.2.4.1 Validation metrics

Performance of habitat suitability models and species distribution models can be assessed at several levels and depends on the type of response provided by the model. A distinction is made between models providing a discrete response (e.g. presence or absence) and models providing a continuous response (e.g. suitability score, density). Models trained with a discrete response variable and providing discrete predictions are easily summarised by means of a confusion matrix. Within this matrix, correct predictions are located on the main diagonal and contrasted with the incorrect predictions off-diagonal. For instance, with the binary presence/absence response variable, correct presence (true positive; TP) and absence (true negative; TN) predictions populate the main diagonal while incorrect presence (false presence; FP) and absence (false absence; FN) predictions are situated off-diagonal (Table 3.3).

Table 3.3: Confusion matrix for calculation of performance measures. *Elements represent true positive (TP) values, false positive (FP) values, false negative (FN) values and true negative (TN) values.*

		Observed	
		Presence	**Absence**
Predicted	**Presence**	TP	FP
	Absence	FN	TN

Common metrics to assess model performance based on this confusion matrix include accuracy (correctly classified instances; CCI), Cohen's kappa statistic (κ), sensitivity (Sn), specificity (Sp), true skill statistic (TSS), odds ratio and the Jaccard index (Fielding and Bell, 1997; Manel *et al.*, 2001). Each of these indices is calculated differently (see Table 3.4) and therefore discusses complementary characteristics of model performance (Mouton *et al.*, 2010). CCI provides the most straightforward calculation of model accuracy (i.e. all correct predictions divided by all predictions), despite being dependent on the class distribution of the response variable within the training data (Manel *et al.*, 2001). Cohen's κ has been suggested as an alternative to CCI as it allows for chance correction, though has received similar criticism.

Table 3.4: Performance metrics used to evaluate model performance based on the confusion matrix in Table 3.3. *CCI represents the correctly classified instances and N is the total number of instances. After Mouton (2008), Goethals (2005) and Fielding and Bell (1997).*

Performance measure	Calculation
CCI	$\frac{TP + TN}{N}$
Misclassification rate	$\frac{FP + FN}{N}$
Sensitivity (Sn)	$\frac{TP}{TP + FN}$
Specificity (Sp)	$\frac{TN}{FP + TN}$
True skill statistic (TSS)	$\frac{TP}{TP + FN} + \frac{TN}{FP + TN} - 1$
Positive predicting power	$\frac{TP}{TP + FP}$
Negative predicting power	$\frac{TN}{FN + TN}$
Odds-ratio	$\frac{TP \cdot FP}{FN \cdot TN}$
Jaccard	$\frac{TP}{TP + FP + FN}$
Cohen's Kappa	$\frac{(TP + TN) - \left(\frac{\left((TP + FN) \cdot (TP + FP) + (FP + TN) \cdot (FN + TN)\right)}{N}\right)}{N - \left(\frac{\left((TP + FN) \cdot (TP + FP) + (FP + TP) \cdot (FN + TN)\right)}{N}\right)}$

The use of these metrics is not fundamentally restricted to categorical response variables, but can be extended to continuous response variables. However, their application requires the transformation of the latter into a set of subjectively defined classes, based on arbitrary thresholds and causing a certain loss of information. Alternatively, the comparison between observations and predictions can be performed in a more quantitative way, including the correlation (r) and determination (r^2) coefficient and the (root) mean squared error ((R)MSE), as described in Table 3.5 (Bennett *et al.*, 2013).

Table 3.5: Performance metrics for models generating continuous output based on predicted (P) and observed (O) values. *N is the total number of instances.*

Performance measure	Calculation
Correlation coefficient (r)	$\frac{\sum(P \cdot O) - \frac{(\sum P \cdot \sum O)}{N}}{\sqrt{\left(\sum P^2 - \frac{(\sum P)^2}{N}\right) \cdot \left(\sum O^2 - \frac{(\sum O)^2}{N}\right)}}$
Determination coefficient (r^2)	$\left(\frac{\sum(P \cdot O) - \frac{(\sum P \cdot \sum O)}{N}}{\sqrt{\left(\sum P^2 - \frac{(\sum P)^2}{N}\right) \cdot \left(\sum O^2 - \frac{(\sum O)^2}{N}\right)}}\right)^2$
Root mean squared error (RMSE)	$\sqrt{\frac{1}{N} \cdot \sum (P - O)^2}$
Mean squared error (MSE)	$\frac{1}{N} \cdot \sum (P - O)^2$

Nevertheless, real-world data often provides a simple binary occurrence statement, while the increased application of ensemble modelling causes a rise in the prediction of probabilities. Discretisation of this score allows model performance assessment via the confusion matrix and classification metrics (Table 3.3 and Table 3.4, respectively), though threshold selection differs among studies and ranges from a fixed threshold at 0.5 over the use of species prevalence to the optimisation of Cohen's kappa (Freeman and Moisen, 2008b). Similarly, assigning numerical values (e.g. translating a presence/absence statement into a 1 or 0 score, respectively) to the original response variable helps the application of the regression metrics (Table 3.5). Alternatively, the receiver operator curve (ROC) represents a commonly applied graphical performance indicator that bridges this discrepancy between observation and prediction data. After applying all possible thresholds, the sensitivity (y-axis) is plotted in relation to the specificity (x-axis) and represents the ROC, which can be summarised in a single indicator by calculating the area under the curve (AUC).

The application of AUC to evaluate model performance is relatively common because of its simplicity, generality and discretisation threshold independency (Phillips *et al.*, 2009; Swets, 1988). Values range between 0 and 1, with 1 indicating perfect discrimination and 0.5 representing similar discrimination as random classification. Drawbacks of this indicator are related with (1) ignoring the model's goodness-of-fit, (2) the AUC being not completely independent of species' presence and (3) model performance in regions that are not practically used is incorporated in the AUC (Lobo *et al.*, 2008). Despite these disadvantages, AUC can still be applied when evaluating predictor importance on final model performance (Barbet-Massin *et al.*, 2014).

3.2.4.2 Validation techniques

Calculation of model performance based on the original training data is inherently biased as the fitted model is familiar with the provided data. Unbiased estimates of model performance are obtained when new and independent data is available, reflecting external model validation. The discrepancy between both validation scores arises and qualitatively reflects the degree of overfitting and the generality of the extracted patterns. When significant differences occur, no reliable predictions will be obtained from the model and the results should be interpreted with caution. Moreover, the development of a single model is highly dependent on the provided data and can therefore be unknowingly biased.

These issues can be partially tackled by increasing the overall data-use efficiency and improved hyperparameter tuning, thereby supporting model regularisation (i.e. increasing model acceptance by reducing its specificity). Within the field of occurrence-based correlative modelling, data is a valuable resource and requires careful consideration prior to removal. By training multiple models with a random subsample of the available data, predictions become an aggregate of a series of individual models and decrease the risk of overfitting. Proper development of multiple models being derived from the same data entails supervised sampling of the data to avoid overly correlated models and can be performed via *k*-fold cross-validation (CV). More specifically, the data is separated into *k* different folds, out of which *k*-1 folds are selected for model training and the remaining fold is used for external model validation (see Figure 3.8). Model training and subsequent validation is repeated *k* times to make sure that every fold has acted once as pseudo-independent validation data. Due to this repetitive model development and increased data-use efficiency, CV is considered to be more trust-worthy than simply splitting the data in a training and validation set (Akratos et al., 2008). A graphical representation of *k*-fold CV is depicted in Figure 3.8 for k_{cv} = 10, though other values for k_{cv} can be used (e.g. k_{cv} = 3 or k_{cv} = 5) depending on the researcher's preference and the overall data availability. Moreover, the value of *k*-fold CV during hyperparameter tuning is illustrated in Box 3.2.

An extreme version of *k*-fold CV is *leave-one-out* CV (LOOCV), where the number of folds is equal to the number of instances minus one ($k_{cv} = N_{inst} - 1$), which is quite common when data is limited. Still, *k*-fold CV decreases the amount of instances (and thus, sample size) for model training due to the exclusion of a single fold, which can be considered an unwanted side-effect. Alternatively, bootstrapping allows to maintain the same number of instances by sampling the original data randomly and allowing certain instances to be present twice or even three times while others remain absent and available for model validation.

Yet, increasing data use efficiency during model training also increases potential bias as there is no completely independent data set to be used for testing the final model. Therefore, it is highly recommended to, prior to repeated model development, extract a subset of the data that is never used for model training (see Figure 3.7). Alternatively, completely new data is collected, reflecting (i) similar environmental conditions, (ii) different environmental conditions or (iii) different geographical regions, depending on the purpose and known limitations of the model.

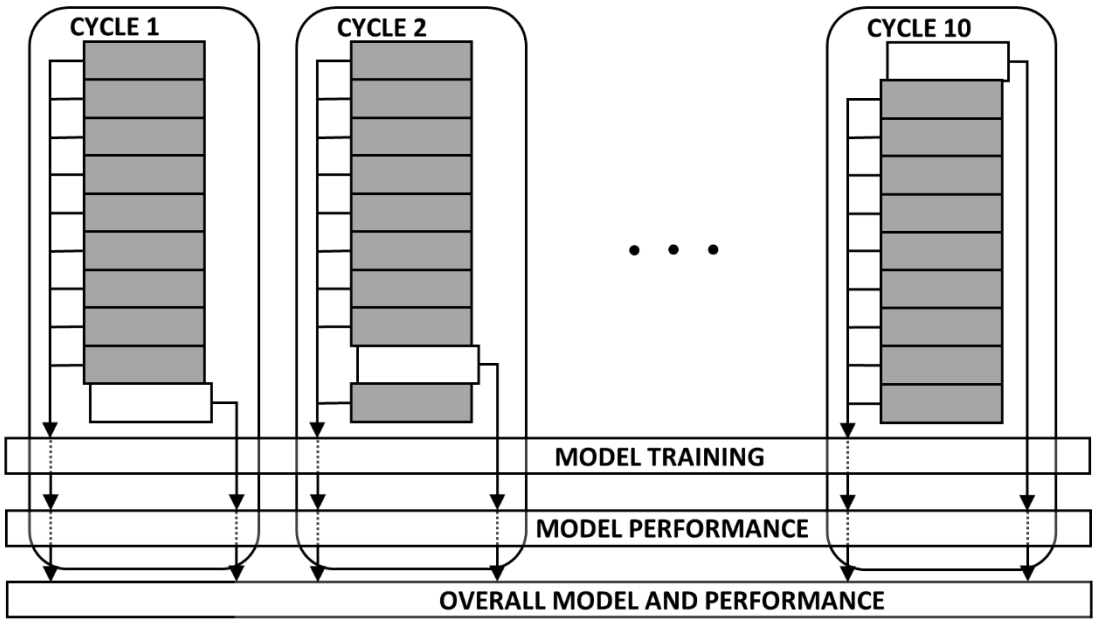

Figure 3.8: Illustration of 10-fold cross-validation during model development. *The initial data set is split in 10 equal-sized folds, out of which a different fold is used for validation during each cycle.*

Box 3.2: Using *k*-fold cross-validation for hyperparameter tuning

Several data-driven modelling techniques are characterised by including a series of hyperparameters (see Section 3.2.3), which require to be defined by the user prior to algorithm application. The selection can be fixed to the default conditions specified by the used software, though the majority of studies benefits from (some kind of) hyperparameter tuning. This can be obtained by repetitive model development and associated performance assessment.

Aside from limiting overfitting and decreasing variance within the final model, *k*-fold cross-validation can also be used for hyperparameter tuning. For each combination of hyperparameter values, *k* different models are developed and assessed as depicted in Figure 3.8. The combination that provides the best performance (see Section 3.2.4.1 for available metrics) is ultimately selected and reported as the implemented hyperparameter settings. Prior to performing such a repetitive assessment of all potential combinations, hyperparameter values need to be defined. This can follow (1) a structured approach with *a priori* definition of all combinations to be tested or (2) an iterative approach based on the results of the previous iteration. A visual representation of using *k*-fold cross-validation for hyperparameter tuning is provided in Figure 3.9.

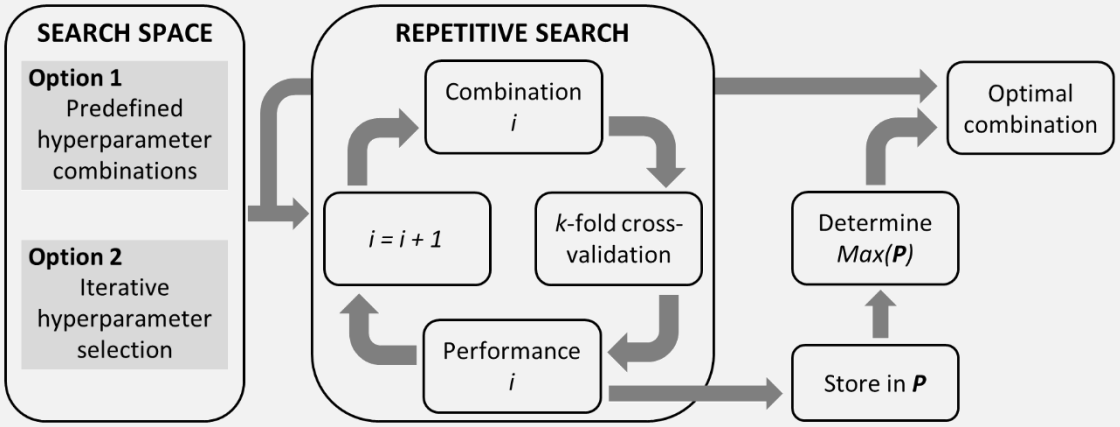

Figure 3.9: Position of* k*-fold cross-validation in hyperparameter tuning. *The extent of the search space can be completely defined (Option 1) or dependent on the observed performance during previous iteration (Option 2). Subsequently,* k *models are developed with a specific combination of the selected hyperparameters and their performance is pooled and evaluated against the performances of all other combinations. Finally, the hyperparameter combination supporting the highest performance is identified and selected.*

3.3 Criticism on data-driven models

First of all, Guisan and Thuiller (2005) mention that, when using observations to predict a species' presence, the obtained model describes the *realised* niche (as part of the *fundamental* niche). Moreover, by using observation data, species are assumed to be in equilibrium with their environment, thereby ignoring tolerance capacity and mobility behaviour (*source-sink dynamics* and *dispersal limitation*, respectively, *sensu* Pulliam (2000)), along with the characteristic disequilibrium displayed by recently introduced species (Gallien *et al.*, 2012). Presence in an unsuitable habitat and absence from a suitable habitat negatively affect the performance of observation-based models (Guisan and Thuiller, 2005; Pulliam, 2000; Sinclair *et al.*, 2010), causing the creation of overly complex models and incorrect distributions of predicted suitable habitats. When willing to describe the fundamental niche, one needs to fall back on autecological experiments and process-based models (Gallien *et al.*, 2010).

Secondly, the pool of existing modelling techniques has increased greatly throughout the last decades and impedes the creation of a useful, concise overview. Whereas in most cases higher diversity is cheered for, here it brings along two important consequences (Guisan and Thuiller, 2005): (i) an increased range of model-specific errors and uncertainties, and (ii) divergence of the modelled response variable. So far, comparative research on both aspects remains insufficient to perform a qualitative comparison of all techniques, illustrating one of the challenges when faced with appropriate model selection (De'ath and Fabricius, 2000). As a partial solution, Araújo and New (2007) suggested to apply ensemble forecasting to combine the predicted responses of several models, resulting in a consensus prediction and a probability range. Despite its promising applicability, ensemble forecasting only provides an end-of-pipe solution and leaves the real causes of the divergence untouched.

Thirdly, scale and resolution vary depending on the type of data (e.g. small-scale nutrient state at high resolution versus large-scale climatic conditions at low resolution) and can lead to the decision of excluding specific variables (Elith and Leathwick, 2009). However, this decreases the transferability of a small-scale model to a larger scale (e.g. outside the originally considered climatic conditions), while up- and downscaling avoids variable exclusion, though introduces errors via data aggregation (reduction of detail) and the assumption of similar conditions (generalisation), respectively. Alternatively, a cascade of models can be developed, starting with global models providing coarse suitability maps, out of which specific areas of interest can be selected for investigation at a smaller, more detailed scale (Roura-Pascual *et al.*, 2009). The development of regional high-resolution models provides a potential bridge between the low-resolution climate change scenarios and high-resolution field observations, though attention should be given to the boundary conformity with the large-scale models.

Finally, overarching these points of criticism, is the limited inclusion of temporal dynamics by HSMs, especially within the framework of forecasting the effects of climate change on habitat suitability. The inherent interactions included in observation-based HSMs are likely to change when climatic conditions differ. For instance, increased atmospheric carbon dioxide causes higher aquatic carbon dioxide concentrations, acidification and elevated temperatures (IPCC, 2014), all of which influence the metabolic processes of organisms in a different way and, by consequence, the prevailing interactions (Gallien *et al.*, 2010). In short, correlative HSMs are a straightforward way of linking habitat conditions to species occurrence, but the underlying assumptions caution the consideration of their response as the one and only truth, additionally highlighting that models are only a mere simplification of reality (Wilson *et al.*, 2011).

3.3.1 Including dispersal dynamics to predict species distributions

Suitable habitats provide potential for a macrophyte to be present, yet natural barriers and limited connectivity decrease dispersion efficiency and thereby impede introduction, establishment and colonisation. Dispersion efficiency greatly depends on the prevailing species pools in the immediate surroundings and the applied dispersion strategy (Galatowitsch, 2006; Sundermann *et al.*, 2011). For instance, Sundermann *et al.* (2011) illustrated that river restoration success largely depends on the surrounding species pools, while indicating that species-specific dispersion rates are limitedly known due to distinct dispersion strategies (e.g. stolons, cloning and root growth).

Dispersal dynamics play a major role in the observation of false absences (i.e. no observation in a suitable habitat) and false presences (i.e. observation in an unsuitable habitat). For instance, false absences are caused by a suboptimal introduction frequency into a suitable habitat. Current absence of the species can be linked with a recent stochastic disappearance or abiotic restoration and is exacerbated by decreased environmental connectivity or a relatively low tendency to disperse (Jiménez-Valverde *et al.*, 2008). Similarly, false presences represent the process of continuous species introduction into a habitat that does not support the development of a viable population and acts as a sink environment (Pulliam, 2000). Both cases illustrate the criticism on the assumption of HSMs that species are in equilibrium with their environment and how this can interfere with consolidating conclusions (Guisan and Thuiller, 2005).

Knowledge of species-specific dispersion rates provides the potential to predict future species distributions and the timeframe needed for a macrophyte to establish and subsequently colonise the identified suitable habitats. Inclusion of species dispersion rates transforms HSMs into species distribution models (SDMs) and can be performed prior to abiotic filtering (Guisan and Rahbek, 2011), although a lack of data impedes its inclusion. This highlights an important field of future study in case short-term restoration via natural succession is aimed for.

3.4 Contribution to the study objectives

Each technique has specific advantages and drawbacks, the latter of which can often be resolved (partially) by technique-specific extensions. Yet, as the number of explanatory variables is expected to be high and only limited expert knowledge is available, it would be unwise to choose FL or BBN. Similarly, GLM development and interpretation is expected to be hampered due to the high number of explanatory variables and potential interactions that require explicit inclusion in the model structure. In addition, the black box behaviour of ANN is hardly resolved via technique-specific extensions, which impedes transparency towards the end user. Finally, DTs suffer from relatively high instability, though this can be reduced by alternative data use methodologies and algorithm application. A specific implementation of repetitive tree development is the random forest technique, which is recommended for further analyses. A tabular overview of technique-specific advantages and drawbacks with respect to the study objectives (see Section 1.2.1) is provided in Table 3.6.

Table 3.6: Drawbacks and advantages of the selected techniques framed within the study objectives.

Technique	Advantages	Drawbacks
DT	- Transparent; - Identifies variable interactions.	- Instability of single tree; - Influenced by data set dimensions.
GLM	- Easy to use	- Assumed distribution; - Explicit inclusion of variable interactions.
ANN	- Tolerates noise and errors; - Identifies variable interactions.	- Black box model; - Lack of guidelines.
FL	- Fuzzy boundaries	- Influenced by data dimensionality; - Data discretisation.
BBN	- Accounts for uncertainties explicitly; - Straightforward error propagation.	- Data discretisation; - Time-intensive rule construction.

Aside from supporting technique selection, this chapter additionally provides a basis for the efficient usage of data within the model development framework (i.e. combining holdout with *k*-fold cross-validation) and the assessment of model performance with the threshold-independent AUC, presence-oriented sensitivity (Sn) and absence-oriented specificity (Sp).

3.5 Conclusion

Both habitat suitability and species distribution models provide a promising approach to support conservation and restoration management in a changing world. A variety of modelling approaches exists with specific advantages and drawbacks, making the selection of a single technique highly subjective. Overall, decision trees are relatively simple techniques allowing for the integration of variable interactions without the need to specify a distribution (GLMs) or a number of hidden layers (ANNs), while the relative recent random forest reports comparatively high performance. The possibility to include ecological knowledge within DTs is relatively limited and requires the use of a more advanced technique like FL or BBN, with the latter showing promising results when combining experimental experiences and expert knowledge. Data becomes, more than ever, a valuable resource and deserves to be treated with care and properly cleaned prior to being mined. To increase data use efficiency and limit model overfitting, cross-validation is to be applied during model development, while performance assessment based on non-transformed observed or predicted response values is most informative. The area under the receiver operating characteristic curve (AUC) represents a single-value and threshold-independent metric, while sensitivity (Sn) and specificity (Sp) provide valuable additional information on model characteristics.

4

Data description and modelling technique

Highlights

- Characterisation of the Limnodata Neerlandica as original data set
- Introduction to conditional random forests and related hyperparameters
- Description of model development (pre-processing, training and evaluation)

Abstract

When aiming to merge several ecosystem services through restoration or artificial creation of wetlands, a profound knowledge of the underlying processes and interactions is crucial. This knowledge can be gathered by relying on field observations to develop habitat suitability models on the one hand and performing autecological experiments to improve fundamental knowledge on behaviour dynamics on the other hand. Data is a major source of information for the development of correlative models, but requires proper identification, characterisation and cleaning prior to being used for pattern extraction. The publicly available Limnodata Neerlandica showed to be very extensive, but prone to high degrees of missing data and elevated levels of faulty or irrelevant information. These issues are tackled with a variety of existing techniques, though trade-offs between information gain and time-related efficiency loss are needed for well-balanced technique selection. Here, data preparation aims at improving the performance of conditional random forests, belonging to the family of decision trees and combining a pre-specified number of individual trees (*ntree*) into a single model to increase response stability. The use of a pseudo-independent test set and five-fold cross-validation supports relatively unbiased performance assessment via the Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity (Sn) and specificity (Sp). Simultaneously, experiments with two *Lemna* spp. under controlled conditions aim at confirming the invasive character and vulnerability to invasion by working at the pre-introduction and post-establishment level. These technicalities create the practical framework and support repeatability of the performed analyses.

4.1 Setting the scene

In previous chapters, literature was consulted to create a framework and to identify key issues for further research within the field of wetland restoration. From Chapter 2, it became clear that attention should be given to both modelling and experimental studies, simultaneously laying the path for all subsequent chapters. As a start, Chapter 3 provided a structured insight into the advantages and drawbacks of five frequently applied modelling techniques, while stipulating that data pre-processing, cross-validation and external validation are essential to obtain a qualitative model. Still, a variety of challenges remains to be tackled, for which a more methodological approach is necessary.

When aiming to merge several ecosystem services through restoration or artificial creation of wetlands, a profound knowledge of the underlying processes and interactions is crucial. This knowledge can be gathered by relying on field observations on the one hand and performing autecological experiments on the other hand. Field observations inherently describe the *realised niche* of the studied species, though only deliver snapshot information on the perceived ecosystem due to the limited inclusion of spatiotemporal dynamics (Araújo and Guisan, 2006). Based on these occurrence data, correlative habitat suitability models (HSMs) are developed to describe or predict species distributions. Yet, with low-quality data undermining many modelling attempts, specific attention is required to tackle the presence of missing data, outliers and redundant variables (Donders *et al.*, 2006; Zuur *et al.*, 2010).

In contrast, controlled conditions provide the opportunity to investigate species-specific traits and dynamics of invasive alien species (IAS), which allows the inference of the mechanisms underlying species dominance (Hofstra *et al.*, 2020; Keddy, 2010). These experiments help developing proactive and reactive management plans by assessing the ability to forecast invasive behaviour and the response of biomass production under a combined management-introduction pressure, respectively. Yet, extrapolation of these results to relevant environmental conditions and spatiotemporal dynamics requires caution.

Within this work, both the realised niche and functional traits are considered to define what constitutes a suitable habitat and how invasion vulnerability can be assessed. Therefore, this chapter is divided into two parts: (1) related to the development of correlative habitat suitability models and (2) related to laboratory experiments under controlled conditions. The first part is characterised by subparts (e.g. data quality, modelling technique), which are discussed in detail throughout this chapter.

4.2 Habitat Suitability Models

The development of correlative habitat suitability models follows a set of crucial steps prior to interpreting and discussing the results (see Chapter 3, Table 3.1). Data is a major source of information in environmental data science, but the quality of publicly available data sets is often questionable (Gibert *et al.*, 2018a; Maldonado *et al.*, 2015). Therefore, it is advisable to identify, characterise and clean the data prior to using the observed occurrences for pattern extraction. Within the following sections, more information is provided on the exploited data, the considered cleaning techniques and the selected modelling procedure.

All calculations, procedures and modelling activities have been developed and performed in RStudio (version 1.1.463 and older), as graphical user interface for R (version 3.6.1 and older) (R Core Team, 2016; RStudio Team, 2015), unless mentioned otherwise. A variety of packages has been used throughout this work and will be introduced when considered appropriate, along with the general packages *reshape*, *ggplot2*, *ggpubr*, *doParallel* and *foreach* to aid data structuring, plotting and parallel computations (Kassambra, 2019; Microsoft Corporation and Weston, 2019a; Microsoft Corporation and Weston, 2019b; Wickham, 2007; Wickham, 2016). In this light, it is worth mentioning that relevant and associated scripts can be found on GitHub (https://github.com/wvechelp/PHDReleases, licensed under MIT licence).

The notation of units follows the guidelines of the National Institute of Standards and Technology (NIST), while averages are reported as *mean ± 1 SD* (with *SD* being the standard deviation) (Barde and Barde, 2012; Thompson and Taylor, 2008). Exceptionally, the standard error of the mean (e) is used instead of the standard deviation (s) in order to stress the accuracy of the observation rather than the variability, which is mentioned clearly when being used.

4.2.1 Dataset characteristics

The Limnodata Neerlandica (Knoben and van der Wal, 2015) contains information on the hydromorphological, physicochemical and biological conditions of Dutch surface water bodies, being collected between 1968 and 2012 throughout the Netherlands. The data set is a collection of observations made by 38 different institutions (see Table A.1) that is owned and made publicly available by the Dutch Foundation of Applied Water Research (STOWA).

Instances are spatially and temporally referenced, with the majority of sampled water body types being lotic waters, lakes, canals and ditches (STOWA, 2001). Within this work, attention was given to the physicochemical and macrophyte data, both of which showing a variable – but overall increasing – number of observations throughout the data collection period.

4.2.1.1 Physicochemical data

In total, 665 variables are listed as being included in the data set, yet the majority of these variables results from highly specific research, causing many variables to not contain any information at all (i.e. 464 variables) or only provide information for a limited number of instances. Indeed, relatively few variables (N_{var} = 14) contain a value for more than 50 % of all instances (N_{inst} = 34 483), with the lowest degree of variable-wise information density being 0.003 % (Figure 4.1A). Consequently, the degree of missing information varies per instance, ranging between 0.5 % and 59.7 % (Figure 4.1B). Within the original physicochemical data, only few instances (N_{inst} = 792; 2.3 %) contain information on more than 20 % of all included variables (N_{var} = 201). An overview of all variables with at least one recorded value (N_{var} = 201) is provided in Appendix, Table A.2. Prior to further analyses, field and laboratory data on conductivity and pH were merged into a single variable (i.e. N_{var} = 199).

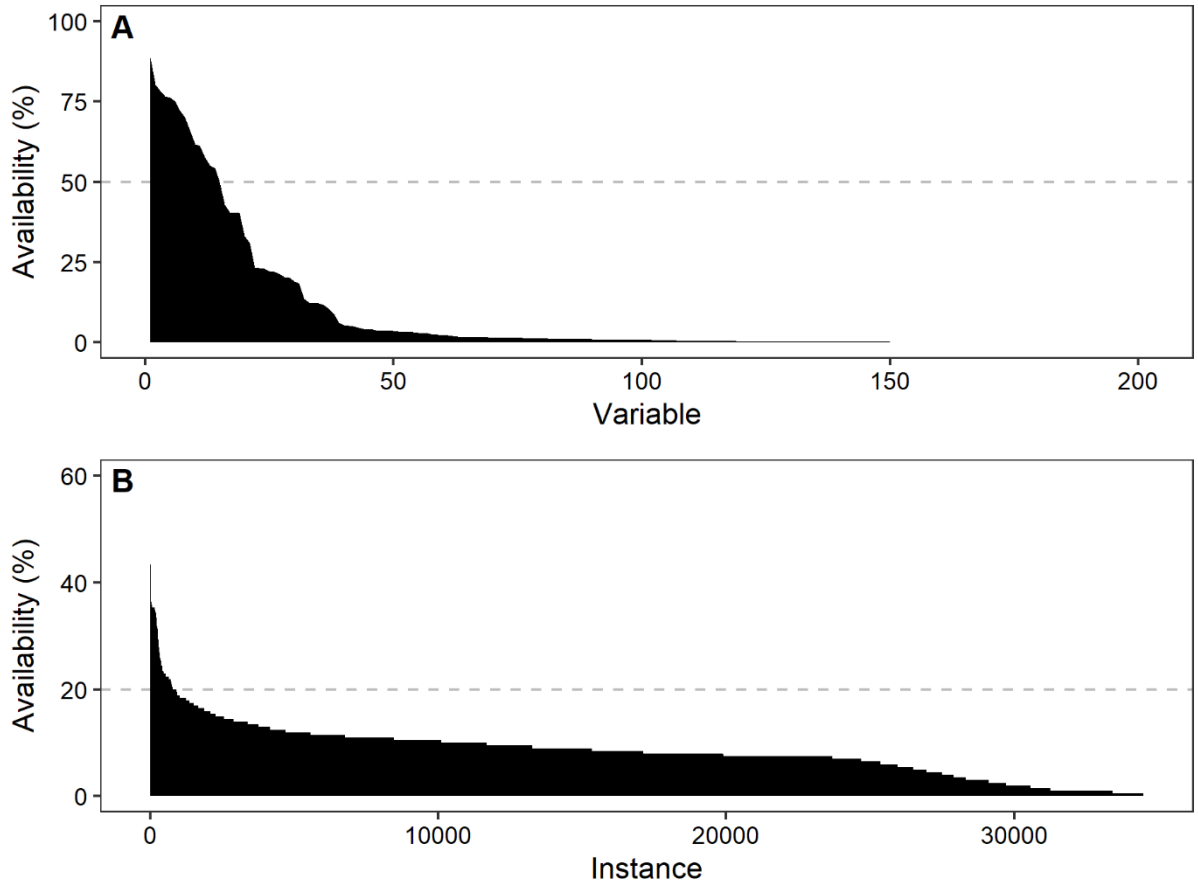

Figure 4.1: Information within the physicochemical dataset. *A: In total, 201 variables contained some information, with the lowest degree being 0.003 % (indiscernible due to the scale of the y axis). Only 14 variables contained information for more than 50 % of all instances (i.e. black surface above the dashed grey line). B: In total, 34 483 instances contained some information, with the lowest degree being 0.5 %. Only 792 instances contained information for more than 20 % of all variables (i.e. black surface above the dashed grey line).*

Unfortunately, detailed information on the applied methodologies, protocols and analytical equipment is lacking within the Limnodata Neerlandica. This impedes extensive quality control within this work and requires the assumption that the majority of the data is collected in a qualitative and standardised manner. Further quality control and data pre-processing are therefore imperative to assess and reduce the amount of noise within the physicochemical data (see Section 4.2.3).

4.2.1.2 Biotic data

Overall, 13 biotic groups are considered in the original data set: Amphibians, Birds, Butterflies, Diatoms, Fish, Macro-algae, Macrofauna, Mammals, Macrophytes, Nematodes, Phytoplankton, Reptiles and Zooplankton. Building on Chapter 2, in-depth attention will be given to the description of the macrophyte data.

Macrophyte occurrence was collected with a variety of techniques, including the Tansley-scale, the Braun-Blanquet method and the basic indication of presence (STOWA, 2001) of which an overview is provided in Table A.3. After removal of misclassified algae, undefined species, hybrid species and ambiguous naming (e.g. only family name), a total of 1148 macrophytes remained. Within the latter, responses suggesting macrophyte presence (e.g. abundance and cover percentage) were replaced by a 'presence' statement, otherwise 'absence' was assumed as to supplement the presence-only data (Elith *et al.*, 2006). Hence, a presence-absence data set was obtained, with the notion that an assigned 'absence' statement does not necessarily reflect a true absence (see Chapter 3) (Anderson and Raza, 2010). Despite the high number of macrophytes remaining in the data set, only a minority (N_{bio} = 23) occurred at more than 10 % of all sites (N_{inst} = 77 200), with the lowest degree being 0.001 % (see Figure 4.2).

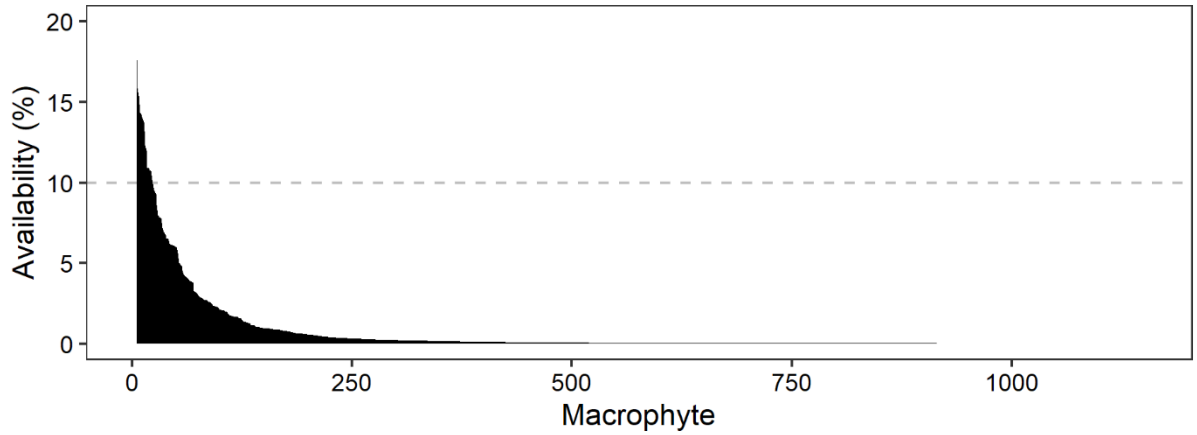

Figure 4.2: Information within the macrophyte dataset. *In total, 1148 macrophytes were observed to be present in at least one location between 1968 and 2012, with the lowest degree being 0.001 %. Only 23 macrophytes were recorded as being present within more than 10 % of all instances (i.e. black surface above the dashed grey line) and thereby represent a clear minority.*

4.2.1.3 Common data

Each instance within the physicochemical and biological data set was characterised by a unique spatiotemporal identifier (STOWA, 2001). Both data sets were reduced to only contain instances with information on the physicochemical and biological situation for each common identifier to avoid a spatial or temporal mismatch between the prevailing abiotic conditions and the observed macrophyte community. Consequently, a significant reduction in data was obtained, with only 4344 instances remaining and simultaneously affecting both the chemical and biotic data sets by reducing the temporal range to the period between 1978 and 2011 (see Figure A.3).

At the physicochemical level, a minor reduction occurred from 199 variables to 174 variables, while at macrophyte level the original 1148 species were reduced to 576 species. During this extensive data selection, the abiotic conditions were assumed to represent the general conditions occurring at that specific location, i.e. that no extreme event had occurred recently.

The resulting combination of physicochemical and macrophyte data were characterised by a geographical distribution throughout the Netherlands (see Figure 4.3, excluding 80 sites that lacked correct georeferencing), which indicates that spatial coverage is not completely uniform. Moreover, a variety of water bodies was sampled, although the majority (N_{inst} = 1729) was not characterised. Additional classes included streams (N_{inst} = 928), brooks (N_{inst} = 926) and canals (N_{inst} = 206).

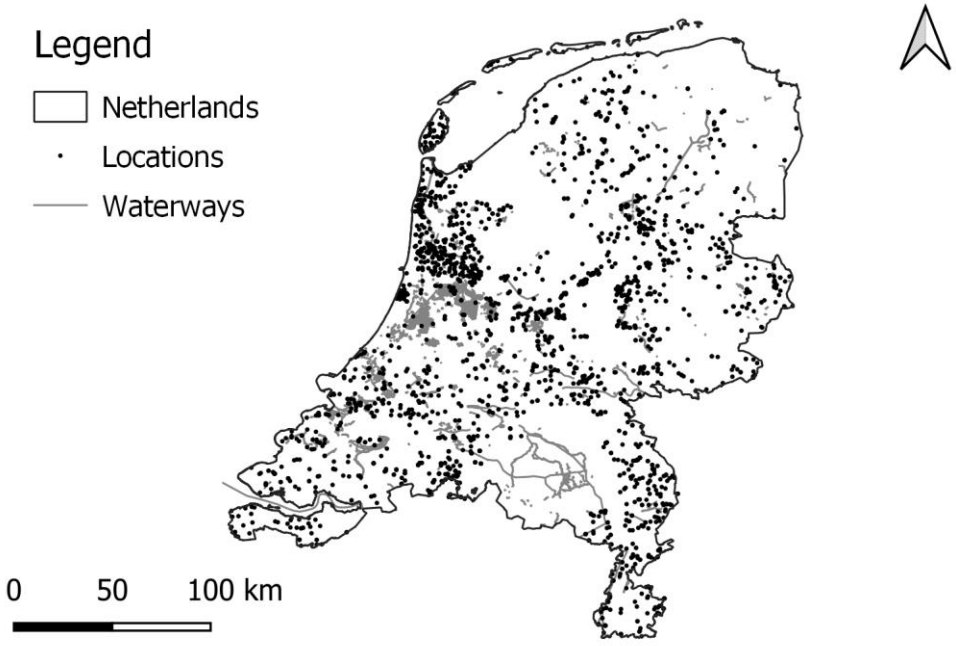

Figure 4.3: Geographical distribution of sampling sites with physicochemical and macrophyte information. *Data is collected between 1978 and 2011 throughout the Netherlands and collected in the Limnodata Neerlandica. A total of 4344 instances were available in the data, but only 4264 were spatially referenced.*

The common data was additionally characterised by a temporal distribution, showing differences between and within years. More specifically, sampling sites were visited and assessed between 1978 and 2011 and showed that during the first years (1978 – 1987), data was collected throughout the year, while this became more restricted in recent years. For instance, samples were initially also collected during the colder months (November – February), while this occurred only sparsely after 1996. This is clearly illustrated by Figure 4.4 and indicates that additional boundary conditions might be necessary for temporal analysis. For instance, to avoid bias when assessing the trends in physicochemical conditions, it can be recommended to only include data from the warmer months (e.g. April – September). Yet, when inferring the realised niche of a macrophyte species, it is assumed appropriate to include all instances, as winter months might represent unsuitable conditions and help in delineating the abiotic habitat that supports the survival and establishment of the considered species.

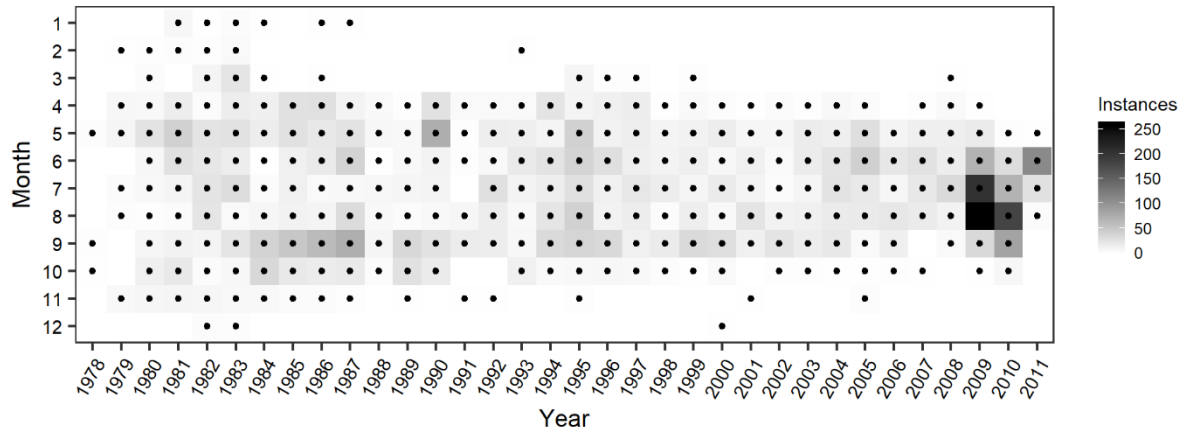

Figure 4.4: Temporal distribution of instances within the combined physicochemical and macrophyte observations. *Instances were collected between 1978 and 2011 throughout the Netherlands and throughout the year. In time, less instances were collected during the winter period (November to February). The dots indicate for which month instances were collected, while the darkness of the tile indicates the number of instances (see also Figure A.3).*

The combined data was characterised by a high degree of missing data points (i.e. 93.7 %), which were all part of the physicochemical data and were distributed differentially over the recorded variables and included instances. For example, only a few variables (N_{var} = 6) contained information for more than 50 % of all instances, with the lowest degree of variable-wise data availability being 0.02 % (Figure 4.5A). At instance-level, only a few locations (N_{inst} = 63) were described by more than 20 % of all recorded variables, with the lowest degree of completeness being 0.57 % (Figure 4.5B). Macrophyte information showed an increase in prevalence for a few species when compared with Figure 4.2, with prevalence ranging between 0.02 % and 40.0 % (Figure 4.5C), though only a minority was recorded at more than 10 % of all locations (N_{bio} = 20).

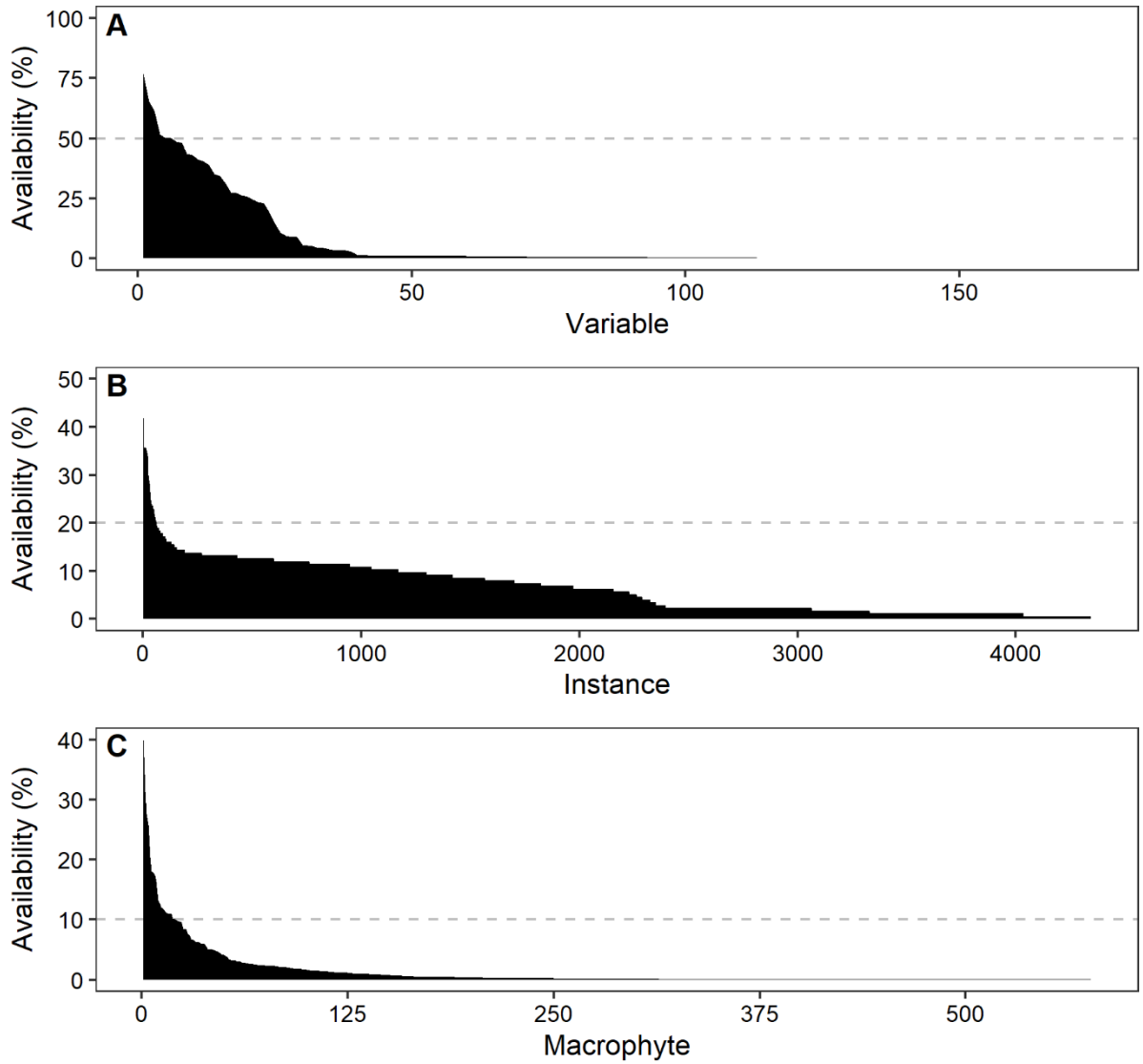

Figure 4.5: Characteristics of the available information within the combined data. *A: Variable-wise information availability; B: Instance-wise information availability within the physicochemical data and C: Macrophyte prevalence depicted as availability.*

The number of missing data points can be reduced by removing incomplete variables and instances from the data set. However, despite providing a reduction in the percentage of missing data, such removal also reduces the number of variables in the data (Appendix A.4, Figure A.4). From this analysis, it is clear that a reduction in missing data can be obtained, though that their presence within the final data set is hard to avoid. For instance, to obtain a reduction from 93.7 % to 50 %, about 154 variables were removed, leaving only 20 variables within the remaining data set. It is therefore considered appropriate to perform data imputation in order to avoid an overly simplified assessed environmental domain. This necessity is caused by the fact that variables have been recorded relatively randomly, as illustrated by a more detailed visualisation of the missing data in Appendix A.4, Figure A.5.

4.2.2 Modelling technique

Model development was conducted via Conditional Random Forests (CRFs) as the resampling strategy and splitting criterion of ordinary random forests (RFs) favour continuous and multinomial variables (Hothorn *et al.*, 2018; Strobl *et al.*, 2007; Strobl *et al.*, 2009b). CRFs belong to the family of decision trees (see Chapter 3) and are an extension of the standard Classification and Regression Trees (CARTs). Throughout this study, CRFs were trained with presence/absence data, though most statements on CRF applicability towards a binary response variable can be extrapolated to multiclass response variables.

4.2.2.1 Principle

A random forest combines a pre-specified number of individual trees (*ntree*) into a single (ensemble) model to increase the stability of the response (Araújo and New, 2007; Breiman, 2001). Each individual tree is trained with a subset of the initial training dataset and, prior to each split, a subset of variables is randomly selected (default $mtry = \sqrt{N_{var}}$, with N_{var} the number of variables). The Gini node impurity ($I(p) = \sum_k p_k \cdot (1 - p_k)$, with k the number of classes and p_k the fraction of instances classified within class k) is calculated to determine the most informative split within the considered variable subspace (i.e. lowest Gini node impurity) (Archer and Kimes, 2008). For each split, a new combination of variables is considered, for which the optimal threshold is sought for within the random subspace. This process of single tree development is repeated multiple times to end up with a series of models consisting of a predefined number of trees (i.e. defined by *ntree*). Because of the random selection of variables for each split, the developed classifiers are only limitedly correlated, allowing to combine (i.e. bagging) the individual responses into an average response (Archer and Kimes, 2008; Strobl *et al.*, 2007). Hence, the final response of the model is determined based on a probability distribution or on a majority vote of all individual trees, with ties assigned randomly (Breiman, 2001; Cutler *et al.*, 2007). The obtained probabilities can subsequently be interpreted as a suitability score.

Advantages of the random forest technique include limited overfitting, robustness towards noise, no need for an a priori assumed variable distribution and the possibility to determine variable importance (Breiman, 2001; Elith and Graham, 2009; Vezza *et al.*, 2015). Yet, also the latter is reported to be flawed within ordinary random forests when variables have different scales or number of categories (Strobl *et al.*, 2007), supporting the decision to develop conditional random forests. Note that throughout this study, the word 'suitability' is used instead of 'probability', as the latter reflects a higher certainty of a species being present, which cannot be reliably obtained when pseudo-absences are included in the model structure (Elith *et al.*, 2005).

4.2.2.2 Model validation and evaluation

Predictions made with CRFs are provided as a probability distribution over the response classes and are situated within the [0 – 1] range, summing to 1 (Hothorn *et al.*, 2018). Discretisation of these probability scores to a presence/absence statement is possible, but requires the selection of a cut-off value, above which a specific instance supports the presence of the considered species. This allows the construction of a confusion matrix, summarising the comparison of observations and predictions into (i) True Positives (TP), (ii) True Negatives (TN), (iii) False positives (FP) and (iv) False Negatives (FN). Based on this matrix, a series of performance metrics can be derived (Chapter 3, Table 3.3 and Table 3.4, respectively).

Despite the claimed robustness of RFs to overfitting, cross-validation is recommended to avoid overly positive performance scores and to increase data use efficiency. During cross-validation, the available data is split into k_{cv} folds, out of which k_{cv}-*1* folds are used to train the model, while the remaining fold provides an estimation of model performance. This is repeated k_{cv} times to make sure every fold has been excluded at least once from the model training step and provides an average performance estimation over all k_{cv} runs (see also Section 3.2.4.2).

Throughout this study, five-fold cross-validation was applied to limit model overfitting and provide information on internal validation. Moreover, model performance was assessed externally by extracting 10 % of each data set as a pseudo-independent test set, while the remaining 90 % represented the basis for creating training sets. The latter was subsampled to obtain a prevalence of 50 % within the final training data set, considering the sensitivity of random forests towards imbalances (Evans and Cushman, 2009; Fox *et al.*, 2017). Ultimately, model performance was reported with AUC and supplemented with sensitivity (Sn) and specificity (Sp) (see Table 3.4) by discretising model output into a presence/absence statement. Threshold selection was determined by minimising the sensitivity-specificity difference (Jiménez-Valverde and Lobo, 2006).

4.2.3 Data preparation and modelling

The combined data (see Section 4.2.1.3) contains information on both the chemical conditions and observed macrophyte community for 3443 instances, yet is still characterised by containing uninformative variables and low-prevalence species. Even though reducing the number of low-informative explanatory variables increases the overall information availability, a high degree of missing values is obtained in the final data set. These missing values only occur within the physicochemical dataset, as absence of information within the biological dataset was considered to represent an absence of the species. While removal of all instances with at least one missing value would cause a high degree of information loss, imputation of the missing values remains a valid alternative as part of the data preparation.

In order to determine the better imputation technique, all physicochemical data was considered to test four different approaches, which is discussed in detail in Chapter 5 (see Figure 4.6). Subsequently, further data pre-processing was applied to eliminate noise both at instance- and at variable level. As these approaches were often linked with the response variable, the combined data was used for these assessments. Identification of the effects of data pre-processing on model performance and computation time is discussed in Chapter 6, while Chapter 7 builds further on these results for species-specific model development (see Figure 4.6).

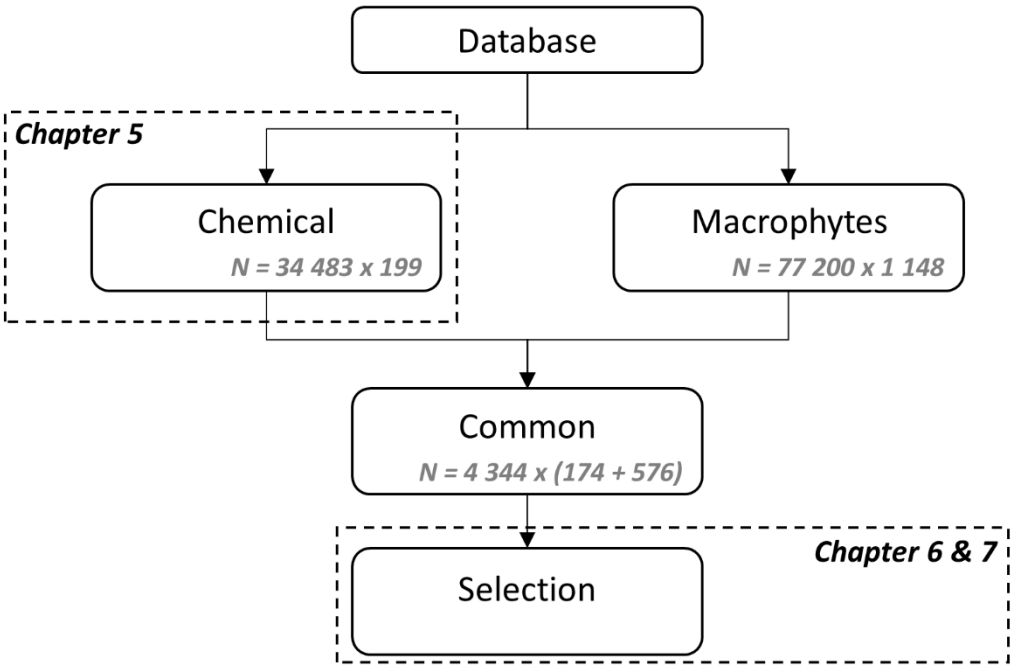

Figure 4.6: Illustration of data use for different chapters. *Each data set is characterised by a certain degree of information, reported as* N = *instances* x *variables. For the combined chemical and macrophyte data a summation of chemical variables and macrophyte species is included, respectively.*

4.2.3.1 Imputation of missing data

Despite being extensive, a high degree of missing data was obtained within the physicochemical data, requiring data reduction to increase the degree of information within the data set. In order to cope appropriately with missing values, Chapter 5 looks into a selection of imputation techniques when imputing artificial missing values.

Characterisation of the data

All physicochemical data (see Figure 4.6) was extracted and contained information on 34 483 unique space-time instances (N_{inst}) and 199 variables (N_{var}), yet being characterised by 90.6 % missing values. Stepwise deletion of variables according to their degree of missing data was followed by determining the total number of complete cases and the accompanying number of data points (i.e. unique instance-variable combinations, $N_{inst,c} \times N_{var,c}$).

The data set containing the highest number of data points without any missing value ($\boldsymbol{D_{opt}}$) was considered as the starting point for the creation of additional data sets, which were fashioned to account for potential variability due to differences in sample size, dimensionality and degree of missing data. First, two additional data sets were created by increasing and decreasing the optimal number of variables ($N_{var,opt}$) with 50 % (Table 4.1). Secondly, 100 %, 75 %, 50 % or 25 % of the instances (N_{inst}) were randomly sampled without replacement, resulting in a total of 12 data sets. Lastly, each data set was subjected to random removal of data points (i.e. equal weights for each variable), representing 1 %, 5 %, 10 %, 25 %, 50 % or 75 % missing data, each being repeated 10 times. Consequently, a total of 720 data sets was considered for imputation. The implementation of this procedure is provided as pseudo-code in Algorithm 4.1.

Table 4.1: Composition of the data sets regarding number of variables and number of instances. *The first complete-case data set contained the highest number of data points. Based on this set, dimensionality for two additional data sets is pre-set during variable removal to act as baseline data (codes 2 and 3). Secondly, three new data sets are derived from the baseline data, with different fractions of instances (codes 4 up to 12).*

Data set code	Variable fraction (%)	Selected instances (%)	Resulting number of variables (N_{var})	Resulting number of instances (N_{inst})	Resulting number of data points
Baseline data					
1	100	100	10	17 264	172 640
2	50	100	5	21 543	107 715
3	150	100	15	3 970	59 550
Derived data sets					
4	100	75	10	12 984	129 840
5	100	50	10	8 632	86 320
6	100	25	10	4 316	43 160
7	50	75	5	16 157	80 785
8	50	50	5	10 771	53 855
9	50	25	5	5 385	26 925
10	150	75	15	2 977	44 655
11	150	50	15	1 985	29 775
12	150	25	15	992	14 880

Algorithm 4.1: Construction of data sets with artificial missing data

Define number of columns $n_{opt,var}$ in $\boldsymbol{D_{opt}}$
Define counter w equal to 1
FOR each element i in [50; 100; 150]
  Identify data set $\boldsymbol{D_{base}}$ with $0.01 \cdot i \cdot n_{opt,var}$ columns
  Define number of instances $n_{base,inst}$ in $\boldsymbol{D_{base}}$
  FOR each element j in [25; 50; 75; 100]
    Randomly sample $0.01 \cdot j \cdot n_{base,inst}$ instances from $\boldsymbol{D_{base}}$
    Store random subset as new data set $\boldsymbol{D_{temp}}$
    Determine number of variables $n_{temp,var}$ in $\boldsymbol{D_{temp}}$
    Determine number of instances $n_{temp,inst}$ in $\boldsymbol{D_{temp}}$
    FOR each element k in [1; 5; 10; 25; 50; 75]
      Define counter z equal to 1
      WHILE $z \leq 10$
        Change seed for different randomisation
        Randomly remove $0.01 \cdot k \cdot n_{temp,var} \cdot n_{temp,inst}$ points from $\boldsymbol{D_{temp}}$
        Store as new data set $\boldsymbol{D_w}$ in list $\boldsymbol{L_{data}}$
        Store information on $\boldsymbol{D_w}$ in list $\boldsymbol{L_{info}}$
        Increase counters w and z with 1
      END while
    END for
  END for
END for

Imputation methods

A variety of imputation techniques exists, ranging from simple variable-specific imputation of the mean over regression-based methods to multivariate model-based approaches. Characteristics of these techniques and subsequent technique selection are discussed in more detail in Chapter 5. Each imputation technique was applied on all 720 data sets identified above in order to assess the applicability of the selected techniques. This was done along a gradient of (i) missing data percentage (f_{MD}), (ii) sample size (N_{inst}) and (iii) dimensionality (N_{var}) to assess how imputation performance can be improved by reducing the degree of missing data, increasing the sample size or increasing dimensionality, respectively.

Evaluation of imputation accuracy

Evaluation was performed by using the normalised root mean squared error (NRMSE), as defined by Equation 4.1, allowing a performance comparison between the different data sets from Table 4.1 and with literature (Stekhoven and Bühlmann, 2012; Troyanskaya *et al.*, 2001). Performance comparison was conducted without considering specific data set configurations (i.e. an overall assessment) and supplemented with the following three cases:

1. Influence of percentage missing data (f_{MD}): specific attention was given to the $\boldsymbol{D_{opt}}$ set (data set 1, Table 4.1) as it was expected to contain the most information and hence to support clearer differences between the imputation techniques.
2. Influence of sample size (N_{inst}): specific attention was given to the $\boldsymbol{D_{opt}}$ data set and derived data sets with lower sample size (data sets 1, 4, 5 and 6, Table 4.1), to provide a link with the previous case.
3. Influence of dimensionality (N_{var}): specific attention was given to the three baseline data sets (data sets 1, 2 and 3, Table 4.1). This case also considered $\boldsymbol{D_{opt}}$ and can be linked with the first case.

$$NRMSE = \sqrt{\frac{\frac{1}{N_{mv}}\sum_{i}^{N_{mv}}(y_i - \hat{y}_i)^2}{\sigma_y^2}} \qquad \text{(Equation 4.1)}$$

With N_{mv} the total number of missing values, y_i the true value, $\hat{y}_i$ the imputed value and σ_y^2 the variance of the true values.

Linear mixed effect models (LMEM) were developed via a backward selection procedure for overall and case-specific performance assessment to infer imputation method significance. Imputation method, degree of missing data, fraction of instances and fraction of variables were considered as (interacting) fixed effects (depending on the considered case), while the imputed data set was included as random effect. Model simplification was performed by stepwise removal of the least significant (interaction) effect, followed by ANOVA testing and (interaction) effect removal if a reduction in complexity (measured via the Akaike Information Criterion, AIC) was obtained. Subsequently, pairwise differences between methods were assessed via post-hoc Tukey tests with Hochberg correction. The *lmerTest* and *multcomp* packages were used for this purpose (Hothorn *et al.*, 2008; Kuznetsova *et al.*, 2017).

Aside from a performance-based evaluation, computation time for each imputation was recorded to qualitatively score each method. This is an often neglected aspect of data imputation and is only limitedly reported in literature as it is subordinate to accuracy (Schmitt *et al.*, 2015). Imputations were run in parallel on two Intel® Xeon® E5620 processors (2.39 GHz and 2.40 GHz), with 6 GB RAM.

4.2.3.2 Data pre-processing

Characterisation of the data

A mismatch between the physicochemical and macrophyte data exists within the Limnodata Neerlandica as data were often collected at different moments in time. Therefore, physicochemical and macrophyte data for space-time combinations that recurred in both data sets were extracted, reflecting the baseline data. Despite being extensive (4344 instances for 174 variables, Figure 4.6), a high degree of missing data was obtained within the physicochemical information (i.e. 93.7 %). Consequently, stepwise variable or instance removal was applied, aiming to reduce the overall degree of missing values. At each step, removal of the variable or instance with the highest positive impact on the overall rate of missing values was performed. Subsequent data set selection and imputation were performed by relying on the results from Chapter 5.

Macrophyte selection was based on the overall number of absolute observations, with at least 100 observations required prior to being included to reduce the original number of macrophyte species (N_{bio} = 576, Figure 4.6). Macrophyte species with lower prevalence can still provide information, yet the limited number of observations creates a highly unbalanced data set, thereby consequently affecting model performance. Remaining macrophytes were subsequently subjected to an additional selection procedure that considered their main life stage habitat, eliminating macrophytes that were more characteristic for bank and terrestrial vegetation. The resulting combined data was used for model development in both Chapter 6 and Chapter 7.

Pre-processing techniques

Due to the specific construction of the random forest algorithm, it was expected that the inclusion of outliers and correlated variables has a limited effect on model performance. However, model regularisation aiming to reduce model complexity by means of reducing incorrect and irrelevant information relies on the trade-off between data and model complexity and, hence, encompasses appropriate instance and variable selection. A variety of pre-processing approaches exists, ranging from simple outlier removal over correlative variable assessment to combinatory algorithm-implemented approaches. The characteristics of these techniques are discussed in more detail in Chapter 6.

Evaluation of pre-processing effects on model performance

The effects of each pre-processing technique on model performance were assessed via the Area Under the Receiver Operating Characteristic Curve (AUC) (see Section 3.2.4.1). Final model evaluation was performed at two levels: (i) using the original external test set and (ii) using the original external test set after pre-processing. By doing so, a performance range can be defined between underperforming models (original test set) on the one hand and overperforming models (pre-processed test set) on the other hand, with the idea that actual model performance lies somewhere in between both results.

Simultaneously, computation time was recorded as it is affected by data pre-processing in two ways: (i) it increases the time needed to prepare the data and (ii) potentially decreases the time needed to develop the individual model. Therefore, computation time was registered for the overall procedure including data preparation and model development as well as for the application of the model training algorithm. The computational capacity was similar as described in Section 4.2.3.1.

4.2.3.3 Model development and habitat suitability assessment

Optimisation of hyperparameters

Optimisation of the selected CRF hyperparameters *ntree* (number of individual models to be developed in the ensemble), *mtry* (number of variables to be considered for each split within the tree), *nsplit* (minimum fraction of instances in a node in order to be considered for splitting) and *nleaf* (minimum fraction of instances in a terminal node in order to be kept) (Hothorn *et al.*, 2018) was conducted based on an iterative, performance-based procedure.

First, an extensive search space was defined by delimiting the ranges of the four hyperparameters and defining the step size between potential values. Ranges differed among hyperparameters (see Table 4.2) and resulted in more than two million possible combinations, out of which sixty combinations were randomly selected to accelerate optimisation (Bergstra and Bengio, 2012). The combination providing the highest AUC score was set as starting point for further parameter tuning.

Table 4.2: Range definition of four hyperparameters. *For each hyperparameter, the baseline, lower and upper limit were defined, as well as the step size. Vector length indicates the resulting number of hyperparameter values. Both* nsplit *and* nleaf *are by default expressed as abolute values, but converted to relative values during optimisation, thereby restricting model complexity.*

Parameter	Baseline	Lower limit	Upper limit	Step size	Vector length
ntree	200	100	1000	10	91
mtry	$\sqrt{N_{Var}}$	2	20	1	19
nsplit	20	0.01	0.5	0.01	50
nleaf	7	0.01	0.25	0.01	25

Secondly, an iterative hyperparameter optimisation procedure was applied by defining a local search space following the hyperparameter-specific range limits as defined in Table 4.3. Identification of the best-performing combination supported the narrowing of the search space by a factor two during the next iteration, yet only when identical hyperparameter values were selected. Iterative parameter optimisation was stopped when the same settings were selected three times or when five iterations were performed (see Algorithm 4.2). This approach does not guarantee finding the global optimum, but helps in identifying a local optimum capable of improving model performance.

Table 4.3: Range definition of four hyperparameters to be used during iterative parameter optimisation. *Within these limits,* x *depicts the frequency of selecting the same settings as providing the highest performance and* y *represents the total number of iterations.*

Parameter	Lower limit	Central	Upper limit	Vector length
ntree	$ntree_{y-1} - \frac{200}{2^x}$	$ntree_{y-1}$	$ntree_{y-1} + \frac{200}{2^x}$	3
mtry	$mtry_{y-1} - \frac{4}{2^x}$	$mtry_{y-1}$	$mtry_{y-1} + \frac{4}{2^x}$	3
nsplit	$nsplit_{y-1} - \frac{0.2}{2^x}$	$nsplit_{y-1}$	$nsplit_{y-1} + \frac{0.2}{2^x}$	3
nleaf	$nleaf_{y-1} - \frac{0.2}{2^x}$	$nleaf_{y-1}$	$nleaf_{y-1} + \frac{0.2}{2^x}$	3

Algorithm 4.2: Iterative hyperparameter optimisation

Develop model *m* with 'starting point settings'
Store 'starting point settings' and AUC in list ***L***
Define iterators *x* and *y*, starting at 0 value
WHILE *x* < 3 and *y* < 5
    Define new search space in list ***S***
    Eliminate settings already occurring within ***L*** from ***S***
    FOR each combination in search space ***S***
        Develop model *m*
        Append list ***L*** with specific settings and AUC from *m*
    END for
    Identify highest AUC in ***L*** and related settings
    IF new settings are the same as 'starting point settings'
        Increase *x* with 1
    ELSE
        Update 'starting point settings' to new settings
    END if
    Increase *y* with 1
END while

Null models, variable importance and partial dependence

Null models were developed for each macrophyte species by randomly permuting the presence/absence statement of the model training data (hence, test data was unaltered), followed by model development with initial hyperparameter settings and external validation with the test data. In total, 1000 null models were developed for each macrophyte and the resulting distribution of AUC values was used to determine the upper 95th percentile (P_{95}). Metric values exceeding this threshold were considered as significantly different from random prediction.

Settings that supported the highest AUC values based on internal cross-validation were subsequently used for final model construction and the determination of variable importance. Variable-specific model improvement ratios (MIRs) were derived for each model and were based on a repetitive permutation-performance assessment scheme (Strobl *et al.*, 2009a). More specifically, the procedure entailed the following steps: (1) a model is trained with the original data, (2) a specific variable of the training data is permuted to break the association with the response variable, (3) a new model is trained with the altered data, (4) a fraction of the data that was not utilised for model training is used to test the new model, (5) the obtained accuracy is compared with the original accuracy and (6) after all individual scores are determined, they are divided by the importance score of the highest-scoring variable. Hence, the obtained MIR score lies between 0 and 1, allowing comparison of relative variable importance among models (Murphy *et al.*, 2010).

Lastly, based on overall importance, five variables were selected for partial dependence plot (PDP) assessment, reflecting the variable's effect on habitat suitability. PDPs were developed by stepwise alteration of the selected predictor along its observed range (minimum-maximum) with the remainder of the training data unaltered, followed by suitability prediction. In total, the PDPs were developed over 21 equidistant values (i.e. 20 breaks) for each of the considered variables.

4.3 Experiments under controlled conditions

Aside from the modelling part, additional attention is given to forecasting the invasive character and the vulnerability to invasion by means of experimental studies. Here, the aim is to work both at pre-introduction and post-establishment level of an invasive alien species by focusing on (1) the applicability of existing trait-based indices to identify an invasive macrophyte and (2) the vulnerability of a system towards invasion, while experiencing an additional management pressure. Experimental conditions varied slightly for these two studies and are therefore introduced separately, within the associated chapter.

Experiments were performed with macrophytes occurring in Belgium, with specific attention towards the selection of a native and an alien species that are preferably phylogenetically close. As floating macrophytes tend to occur in more eutrophic conditions (Bakker *et al.*, 2013; Zhang *et al.*, 2017), potential test species were narrowed down to this subcategory. Among these floating macrophytes, *Lemna minuta* is known for originating from North and South America and having reached a widespread status throughout Europe (Hussner, 2012). In Belgium, *L. minuta* has been observed since 1972 (https://waarnemingen.be/) and is considered to be 'widespread with a moderate impact' (http://ias.biodiversity.be/), while in the Netherlands it has only been observed since 1989 (https://waarnemingen.nl/) and included in the Limnodata Neerlandica since 1990. Four other *Lemna* spp. occur throughout Belgium, being *L. minor*, *L. gibba*, *L. trisulca* and *L. turionifera* (Lambinon *et al.*, 1998; Van Landuyt, 2007). All *Lemna* spp. are characterised by a single root and mostly vegetative reproduction, although sexual reproduction via flowering has been reported too. In order to contrast the performance of the alien *L. minuta* with a native species, *L. minor* was selected as it is a reference species for ecotoxicological studies (OECD, 2006). Consequently, specific guidelines for testing under controlled conditions have been issued, which provides a standardised framework with respect to growth medium, light conditions and potential growth rate.

Both *Lemna* spp. are frequently occurring and well-known for their high reproduction rate, protein content and manipulability (Gérard and Triest, 2014; Yu *et al.*, 2014). Consequently, their potential in treating eutrophic (waste)waters in combination with biomass production has been explored for decades (e.g. Culley and Epps (1973), Hammouda *et al.* (1995), Oron *et al.* (1988) and Yu *et al.* (2014)). On the other hand, their presence in natural systems is frequently characterised by dense mats that decrease light penetration and oxygen concentration, thereby negatively affecting aquatic life underneath these mats (Janes *et al.*, 1996; Janse and Van Puijenbroek, 1998). Hence, their relative similarity and controversial effects on ecosystem structure and functioning provide acceptable arguments to test the applicability of functional traits and the consequences of partial eradication.

5

Imputation methods for missing environmental data[3]

Highlights

- Random forest-based method generally performs best
- Least-squares is valid alternative when computation time is limited
- Data dimensionality has a clearer effect on accuracy than sample size

[3] This chapter is redrafted from Van Echelpoel, W.; Bruneel, S. and Goethals, P. L. M. (submitted) Empirical evaluation of four data imputation methods for incomplete environmental data with varying levels of available information

And additionally based on Van Echelpoel, W. and Goethals, P. L. M. (2018) Variable importance for sustaining macrophyte presence via random forests: data imputation and model settings. *Scientific Reports* 8, 14557, doi: 10.1038/s41598-018-32966-2.

Abstract

A recurrent issue within environmental data sets that impedes appropriate data exploration, analysis and evaluation is the presence of missing data (MD). Existing techniques avoid unnecessary information loss by exploiting available information to impute MD, though individual accuracies differ. Four techniques were selected for comparison of accuracy and required computation time: mean, least square (*ls*) regression, *k* nearest neighbours (*kNN*) and the ensemble-based *missForest* algorithm. Data points were artificially removed from twelve complete data sets (combining three levels of data dimensionality and four levels of sample size) with six different rates of MD, being repeated ten times. Results showed that mean imputation provided stable imputation performance along the MD gradient with an average normalised root mean squared error (NRMSE) of 0.96 ± 0.04, while *ls* and *missForest* provided rather similar performance (0.5 ± 0.3 versus 0.5 ± 0.2, respectively). Higher rates of MD caused an undisputable decrease in performance, except when mean imputation was applied. Simultaneously, computation time increased for *ls* and *kNN*, decreased for *missForest* and remained relatively stable for mean. Sample size affected performance only limitedly, while clearly affecting computation time for *ls*, *kNN* and *missForest*. In contrast, increased data dimensionality positively affected performance, while confirming that computation time was mostly influenced by the total number of data points. Further optimisation of both *kNN* and *missForest* showed a similar increase in performance (Δ_{NRMSE} = -0.05 ± 0.05), confirming that the latter indeed provides better imputation performance than more conventional techniques. In short, the ensemble-based *missForest* algorithm outperformed mean, least squares and *k* nearest neighbour imputation, though the latter two remain valid alternatives at low rates of missing data.

5.1 Setting the scene

Gathering information, improving knowledge and steering decisions all greatly rely on data collection and availability. Yet, many data sets are plagued with a certain degree of missing data as, in practice, data is potentially lost, erroneously recorded or absent due to electronic malfunctioning or non-response (García-Laencina *et al.*, 2010; Giustarini *et al.*, 2016). Missing data is common within the field of environmental monitoring and assessment affecting both descriptive and correlative analyses. For instance, Srebotnjak *et al.* (2012) pointed out that missing data hampered proper water quality index computation, while Chandramouli *et al.* (2007) acknowledged that missing microbiological data impeded accurate human health risk assessment. Moreover, when reviewing watershed-wide water quality evaluation, Olsen *et al.* (2012) observed that 10 out of 49 studies (20 %) reported missing data, with only 1 study reporting the actual degree of missing values. This mismatch between data quality and subsequent data analyses partially underlies reduced efficiency due to the loss of valuable information and a lack of specific guidelines (Giustarini *et al.*, 2016; Liew *et al.*, 2011).

For years, data sets were reduced to contain only complete cases, thereby impeding proper estimation of population parameters, limiting data analysis power and introducing bias (Little and Rubin, 2002; Penone *et al.*, 2014). These complete-case analyses assume that the reduced data set represents a perfect subsample of the population, i.e. a *missing completely at random* (MCAR) mechanism (Little and Rubin, 2002), although most data sets follow the *missing at random* (MAR) or the *not missing at random* (NMAR) mechanism. The latter occurs when data is missing because of its value (e.g. a concentration below detection limit, sensor malfunctioning during a heatwave), while no link can be found with any other variable. In between MCAR and NMAR, the MAR mechanism is characterised by the possibility of estimating missing values based on other variables' values (Little and Rubin, 2002). The increased awareness on the complete-case analysis being acceptable up to only 5 % missing data (García-Laencina *et al.*, 2010) in combination with abovementioned mechanisms, steered the development of imputation techniques.

One of the simplest imputation approaches is based on variable-specific statistics (e.g. mean, median, mode) and represents a popular approach due to fair performance (Celton *et al.*, 2010; Schmitt *et al.*, 2015), despite ignoring the inherent associations among the included variables (Liew *et al.*, 2011). In contrast, a variety of imputation methods do acknowledge these underlying associations, including regression-based methods, Bayesian principal component analysis (bPCA), singular value decomposition (SVD), *k* nearest neighbours (kNN), fuzzy k-means, artificial neural networks (ANN), random forests and model-based approaches (Bø *et al.*, 2004; Brock *et al.*, 2008; Celton *et al.*, 2010; Chandramouli *et al.*, 2007; Luengo *et al.*, 2012; Zhang *et al.*, 2008).

For instance, Bø *et al.* (2004) applied least-squares regression to impute microarray data and concluded that it was simpler and more accurate than *kNN*, despite increasing multicollinearity (García-Laencina *et al.*, 2010). In *kNN*, a pre-specified number of neighbours (k_{nn}) acts as donor for the missing value, representing a tuneable similarity-based imputation. Identifying neighbours is computationally slower compared to statistic- or regression-based imputation and neglects negative correlations, yet often supports higher performances, except when confronted with more advanced techniques (Penone *et al.*, 2014; Schmitt *et al.*, 2015; Waljee *et al.*, 2013). For instance, Stekhoven and Bühlmann (2012) introduced the random forest-based *missForest* algorithm and acknowledged its value for missing data imputation, though optimisation and overall computation time provide a practical trade-off during method selection (Shah *et al.*, 2014; Waljee *et al.*, 2013). A summary of advantages and disadvantages of the mentioned techniques is provided in Table 5.1.

Table 5.1: Advantages and disadvantages of a selection of imputation methods. *Methods include a generally known method (*mean*), a regression-based method (least squares;* ls*), a similarity-based method with limited flexibility (*kNN*) and a random forest-based method with high flexibility (*missForest*).*

Method	Advantages	Disadvantages
Mean	- Simple - Frequently used	- Neglects covariance - Narrows variable distribution - Underestimates variance
Least squares	- Simple - Maintains covariance structure	- Increases multi-collinearity - Does not include local variability - Requires predefined distribution
kNN	- Similarity-based - Can be optimised	- Has to recalculate all distances for each missing value (computation time) - Fixed number of neighbours - Does not include negative correlations
missForest	- Correlation-based - Can be optimised - Flexible related to duplicates	- Optimisation can be cumbersome - Random selection can affect result - Potentially high computation times

Abovementioned techniques share the advantage of single-value imputation, producing a data set that can be used directly for further analysis, yet ignoring the inherent uncertainty of the imputed value. Indeed, a confidence interval can be assigned to each imputed value reflecting the value's potential distribution. Multiple-value imputation methods assume a distribution of the missing value, out of which *m* single values are randomly selected, resulting in *m* new data sets and *m* individual analyses, which are subsequently pooled to obtain an overall evaluation (Faris *et al.*, 2002). Approaches to multiple-value imputation include multivariate normal imputation (MVNI), assuming a multivariate normal distribution, and fully conditional specification (FCS) (Lee and Carlin, 2010), which includes the multiple imputation via chained equations (MICE) method for which relatively high performances have been obtained (Schmitt *et al.*, 2015).

Usage and comparison of imputation methods (both single- and multiple-value) is common within the field of bioinformatics (e.g. microarray data), medicine and marketing (Bø *et al.*, 2004; Lee and Carlin, 2010; Nogueira *et al.*, 2007; Shrive *et al.*, 2006; Troyanskaya *et al.*, 2001), though remains limited within purely environmental data analysis. Moreover, comparisons lag behind as new techniques are constantly being developed while previous methods have not yet been sufficiently applied, described and tested.

Within this chapter, four single-value imputation methods are selected to deal with missing environmental data: the mean (the 'popular' approach), iterative least squares (regression-based), *k* nearest neighbours (similarity-based) and random forests (iterative correlation-based). The aim is to elucidate the differences between the imputation techniques at performance level, supplemented with required computation time. To do so, the imputation error will be assessed along a gradient of (i) missing data percentage (f_{MD}), (ii) sample size (N_{inst}) and (iii) dimensionality (N_{var}). For each technique, it is expected that imputation accuracy is positively affected by (i) decreasing f_{MD}, (ii) increasing N_{inst} and (iii) increasing N_{var}.

Based on abovementioned literature and technique-specific characteristics, it is hypothesised that performance-based ranking will provide the following result: random forests > *k* nearest neighbours > iterative least squares > mean imputation. By tackling this issue, a partial answer to RQ2.1 is formulated, with respect to objective 2.1 as identified in Chapter 1. Hence, this chapter concludes with a statement on the suggested imputation technique and how accuracy can be improved by changing the degree of missing data, sample size or data dimensionality.

5.2 Materials and methods

5.2.1 Characterisation of the data and evaluation methods

The analyses performed in this chapter made use of the physicochemical data within the Limnodata Neerlandica, as described in Chapter 4 (see Section 4.2.3.1). In general, the provided data was used as a basis to develop 720 different data sets, which vary at the level of sample size (N_{inst}), dimensionality (N_{var}) and degree of missing data (f_{MD}). Techniques were compared at performance level by means of the normalised root mean squared error (NRMSE) and supplemented with evaluation of the computation time. More detailed information can be found in Chapter 4 and in Appendix B.1.

5.2.2 Imputation techniques

Four single-value imputation techniques were selected for this study: (i) mean imputation (*mean*), (ii) iterative least squares (*ls*), (iii) *k*-nearest neighbours (*kNN*), and (iv) a random forest-based algorithm *missForest* (*mF*). All techniques were initially applied with their default settings and, if applicable, tested for potential optimisation via (i) inclusion of additional information and (ii) iterative hyperparameter setting.

Imputation of the mean is the simplest approach and has been applied at instance- and variable level. Despite its application within microarray research (Troyanskaya *et al.*, 2001), instance-wise imputation of the mean is not considered appropriate with environmental data, hence a variable-wise imputation is applied. Imputation is performed via the *Hmisc* package (Harrel, 2018).

The iterative least squares method assumes an underlying linear relationship among the variables within the data set, thereby supporting its successful application within the field of microarray analysis (Bø *et al.*, 2004; Brock *et al.*, 2008; Zhang *et al.*, 2008) and its potential within the field of environmental data. Imputation is based on the description provided by Bø *et al.* (2004), starting with the imputation of the variable-wise mean, after which the covariance matrices ($\boldsymbol{S}$) are determined and used to solve Equation 5.1. Following the first imputation, means and covariance matrices are updated and a new imputation value is determined until convergence. Here, maximally 10 iterations were run as additional iterations resulted in relatively minor changes within the covariance matrix.

$$\widehat{y_i} = \overline{y_i} + \boldsymbol{S}_{y_i x}\boldsymbol{S}_{xx}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}) \qquad \text{(Equation 5.1)}$$

With $\widehat{y_i}$ the estimated value (to be imputed), $\overline{y_i}$ the average value over $y_i, \dots, y_n$, $\boldsymbol{S}_{y_i x}$ the covariance matrix (vector) between the variable with missing value and the remaining variables, $\boldsymbol{S}_{\boldsymbol{xx}}$ the covariance matrix among the remaining variables, $\boldsymbol{x} = [x_1, x_2, \dots, x_k]^T$ the variables' values for the considered instance and $\overline{\boldsymbol{x}} = [\overline{x_1}, \overline{x_2}, \dots, \overline{x_k}]^T$ the variables' average values.

The *kNN* approach is a distance-based method and uses the information of the k_{nn} closest neighbours of the instance with a missing value. Subsequently, the mean (or median) of these k_{nn} neighbours is used to replace the missing value, optionally weighted for the neighbours' distance from the instance. Within this study, imputation is based on the Gower distance and the distance-weighted average of k_{nn} neighbours. At first, the default value of k_{nn} = 5 is considered for imputation, followed by an assessment of how NRMSE-based optimisation of k_{nn} can improve imputation performance. This optimisation is conducted for each combination in Table 4.1 at six levels of missing data and two repetitions (i.e. *N* = 144, see Appendix B.2). Imputation via *kNN* is applied via the *VIM* package (Kowarik and Templ, 2016).

Lastly, the *missForest* algorithm was introduced by Stekhoven and Bühlmann (2012) and relies on the random forest technique (see also Box 3.1). This technique belongs to the data-driven supervised machine learning classification and regression trees (CARTs) and has been reported to outperform more traditional methods as it creates an ensemble of independent trees rather than a single tree (Stekhoven and Bühlmann, 2012; Waljee *et al.*, 2013). As such, it can be considered as a multiple-value imputation technique, although only a single imputed data set is obtained.

Imputation via random forest works iteratively, comparing each imputed value with its previous value and combining this in an overall difference. Baseline imputation is performed via variable-wise mean imputation, while the stopping criterion is defined as the moment when the calculated difference starts to increase again, as defined by Equation 5.2 for continuous variables (see Stekhoven and Bühlmann (2012) for discrete variables). Alternatively, the number of iterations can be defined *a priori* to avoid non-convergence errors.

$$\Delta_{\boldsymbol{X}} = \frac{\sum_{j=1}^{k}\left(\boldsymbol{D}_{new}^{imp} - \boldsymbol{D}_{old}^{imp}\right)^2}{\sum_{j=1}^{k}\left(\boldsymbol{D}_{new}^{imp}\right)^2} \qquad \text{(Equation 5.2)}$$

With $\boldsymbol{X}$ the set of *k* continuous variables and $\boldsymbol{D}$ the data matrix.

Within this chapter, random data sampling within *missForest* was performed without replacement and three hyperparameters were selected for optimisation: *ntree*, *mtry* and *nodesize*. At first, hyperparameters were set at their default values (i.e. $ntree = 100$, $mtry = \sqrt{N_{var}}$ and $nodesize = 1$), with maximally 10 iterations. Subsequently, these hyperparameters were iteratively altered for each combination mentioned in Table 4.1 at all six levels of missing data and two repetitions (i.e. *N* = 144, see Appendix, Section B.2.2), followed by an analysis of the difference in performance. The *missForest* algorithm was implemented as part of the *missForest* package (Stekhoven, 2013).

5.3 Results

All imputation methods obtained in at least 94 % of the cases a NRMSE value lower than 1. Ranges differed, with *ls* representing the narrowest range (0.03 up to 2.36) and *kNN* the widest range (0.05 up to 3.73). Both *mean* and *mF* scored in between, ranging from 0.89 up to 4.10 and from 0.06 up to 3.63, respectively (Figure 5.1). Best overall performance was obtained by *mF* (0.45 ± 0.27) and *ls* (0.47 ± 0.26), followed by *kNN* (0.53 ± 0.31) and reflecting a clear difference from *mean* (0.97 ± 0.12).

Indeed, higher NRMSE values were observed for *mean*, represented by scores of *ls*, *kNN* and *mF* being mostly situated underneath the agreement line (Figure 5.1). Moreover, the majority of *kNN* results are positioned above the *mF*-based agreement line and, vice versa, the majority of *mF* results are situated below the *kNN*-based agreement line (Figure 5.1). No clear difference is observed between the results for *ls* and *mF*, as indicated by NRMSE values at both sides of the *ls*- and *mF*-based agreement lines (Figure 5.1). These observations are confirmed by the adjusted Tukey test, showing that *mean* performed significantly worse than *ls*, *kNN* and *mF* ($p < 0.001$ for all pairwise tests), while differences among the latter three methods were non-significant ($p > 0.05$).

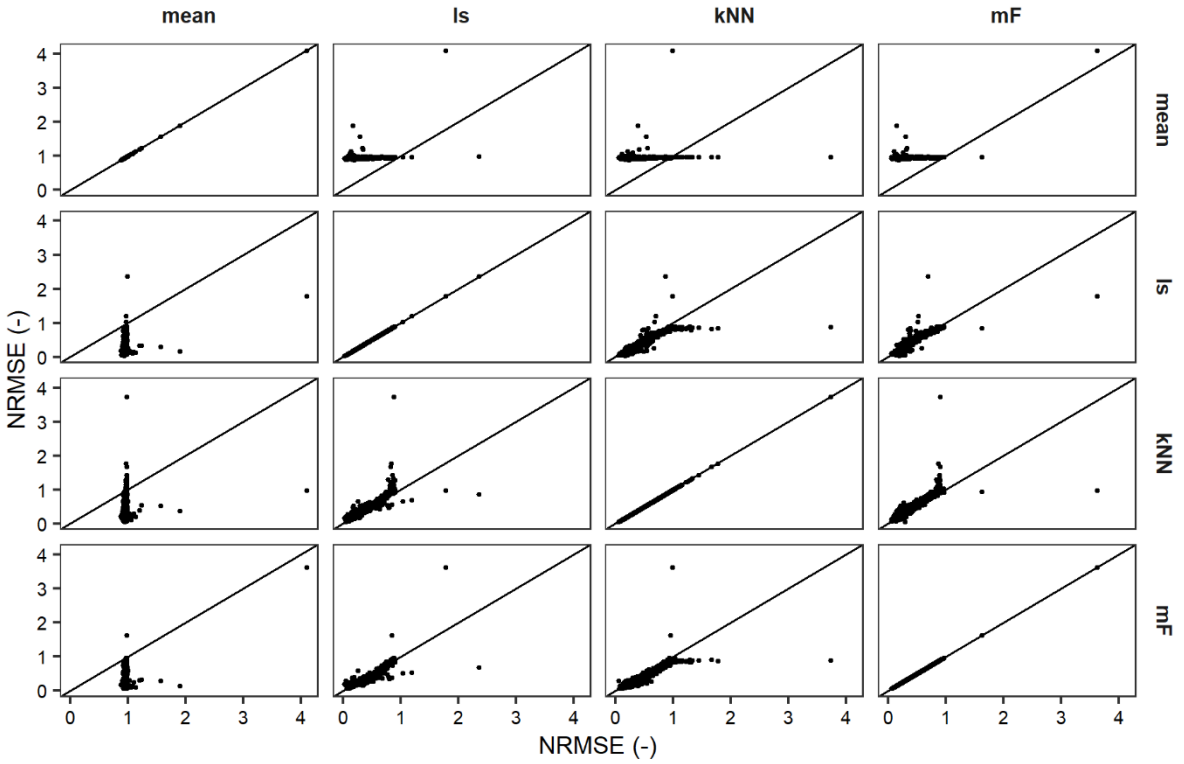

Figure 5.1: General overview of the NRMSE scores for each imputation approach, conditional to the other methods. *To improve visualisation, the y-axis was chosen to be similar to the x-axis range. Values below the agreement line indicate better performance of the method on the y-axis, while values above the agreement line indicate better performance of the method on the x-axis. Methods:* mean*: mean imputation;* ls*: iterative least squares;* kNN*: k nearest neighbours and* mF*: the missForest algorithm. NRMSE: Normalised Root Mean Squared Error.*

In the following sections, more specific results are presented, focusing on the methods' variability in performance and required computation time for (i) a fixed number of both variables and instances (i.e. $\boldsymbol{D_{opt}}$), (ii) a varying number of instances, given a fixed number of variables ($N_{var,opt}$ and flexible N_{inst}) and (iii) a variety in dimensionality (flexible N_{var}). A detailed overview of performance scores can be found in Table B.3. Moreover, in order to support the obtained NRMSE scores with a variable- and technique-specific accuracy assessment, two case studies are provided in Appendix B.4: (i) a small data set (5 variables, 5385 instances) with 1 % missing data and (ii) the optimal data set (10 variables, 17 264 instances) with 50 % missing data. The latter is based on the description of the common data in Section 4.2.1.3. Based on these results, *mF* seemed to perform best for imputing both extensive and confined variables, while *kNN* and *ls* showed to be less applicable, respectively.

5.3.1 Baseline performance at fixed sample size and dimensionality

Highest imputation performance was hypothesised for the lowest amount of missing data, while an increasing degree of missing data (MD) was expected to inflate the imputation error. Separation of the results for imputing $\boldsymbol{D_{opt}}$ conditionally to the degree of missing data clearly supported this hypothesis, with performance of *ls*, *kNN* and *mF* decreasing along an increase in missing values (Figure 5.2). Only *mean* provided consistent imputation performance regardless of f_{MD}.

Based on the saturated mixed model, a lower effect of missing data on *mF* is inferred when compared to *kNN* ($\beta_{mF:MD}$ = 0.859 versus $\beta_{kNN:MD}$ = 1.080), while the discrepancy with *ls* is less clear ($\beta_{ls:MD}$ = 0.822), though significant (p = 0.02). Indeed, *kNN* performance was 0.19 ± 0.05 at 1 % MD, going up to 1.05 ± 0.06 at 75 % MD, while for *mF* this was only 0.16 ± 0.05 and 0.86 ± 0.02 (see Table B.3), respectively. In contrast, no significant difference was observed between *ls* and *kNN*. Moreover, *mean* was unaffected by the degree of missing data (Figure 5.2) and provided an overall stable, yet relatively low, imputation performance of 0.966 ± 0.003 (N = 60), thereby performing significantly worse than *ls*, *kNN* and *mF* (all $p < 0.001$). In addition, *mF* performed significantly better than *kNN* and significantly outperformed *ls* when missing data was at least 20 % (all $p < 0.05$).

Contrasting its performance, *mF* required long computation times, being up to 40 times higher compared to *ls* (e.g. 2100 ± 400 s versus 80 ± 20 s, respectively, with 1 % missing data) and even more when compared with *mean* (0.005 ± 0.008 s, with 1 % missing data). As missing data increased, a decrease in computation time was observed for *mF* (Figure 5.2). Simultaneously, *kNN* showed an increase in computation time, arising to a maximum at 50 % missing values (807 ± 3 s), while *mean* provided short computation times regardless of the degree of missing values (overall 0.006 ± 0.008 s).

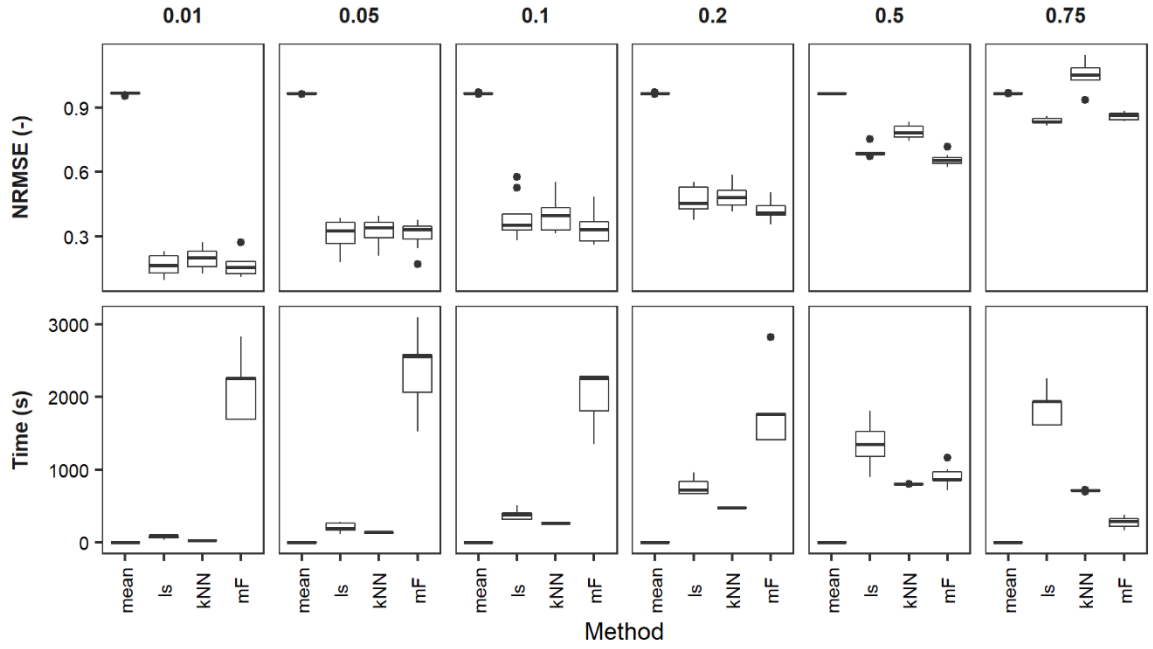

Figure 5.2: General performance of four imputation methods, as determined for the maximum number of data points. *The top row (NRMSE) represents all performance values, while the second row represents the computation time needed for the imputation. Columns represent the different degrees of missing data used. Data of 10 repetitions (identical number of data points, different missing values) are reported. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range. Dots represent the values outside the range of the whiskers. Methods:* mean*: mean imputation;* ls*: iterative least squares;* kNN*: k nearest neighbours and* mF*: the missForest algorithm. NRMSE: Normalised Root Mean Squared Error.*

5.3.2 Sample size variability

Imputation performance was expected to decline with decreasing sample size, vice versa providing higher performance when more data is available. Indeed, imputation error decreased slightly when sample size increased (Figure 5.3), having a relatively higher effect on *kNN* than on *mF* based on the interaction coefficients ($\beta_{kNN:inst}$ = -0.179 versus $\beta_{mF:inst}$ = -0.070, respectively), with *ls* experiencing a similar effect as *kNN* ($\beta_{ls:inst}$ = -0.174). This discrepancy between *mF* and *kNN* created a significant difference in overall performance (p < 0.001) in favour of *mF*, while *ls* and *kNN* illustrated similar performance. Nevertheless, in contrast to the aforementioned significant differences between *kNN* and *mF* at maximum sample size (all p-values < 0.05), both methods performed similarly when imputing smaller-sized data sets with maximally 10 % missing values (most p > 0.05). Likewise, no significant differences between *ls* and *mF* could be observed when maximally 10 % of the data is missing, even at maximum sample size.

At elevated degrees of missing values (≥ 20 %), no clear uniform results were obtained, suggesting a potential dependency on which instances were either in- or excluded. Similarly, the effect of sample size at 1 % missing data remained ambiguous, while at 75 % missing data *kNN* was clearly outperformed by *ls* and *mF* ($p < 0.001$), providing almost similar performance as *mean*.

Only *mean* provided stable and low computation times regardless of the degree of missing data or the number of instances (overall 0.005 ± 0.007 s). On the other hand, *mF* and *kNN* required more time when more instances were provided (e.g. 2100 ± 400 s versus 160 ± 50 s and 25.1 ± 0.2 s versus 2.20 ± 0.09 s, respectively, for 100 % and 25 % of $N_{inst,opt}$, respectively, at 1 % missing data), along with a general increase in computation time for *kNN* when more data was missing and a decrease in computation time for *mF* when more than 20 % of the data was missing (Figure 5.3), reflecting the pattern as observed in Figure 5.2.

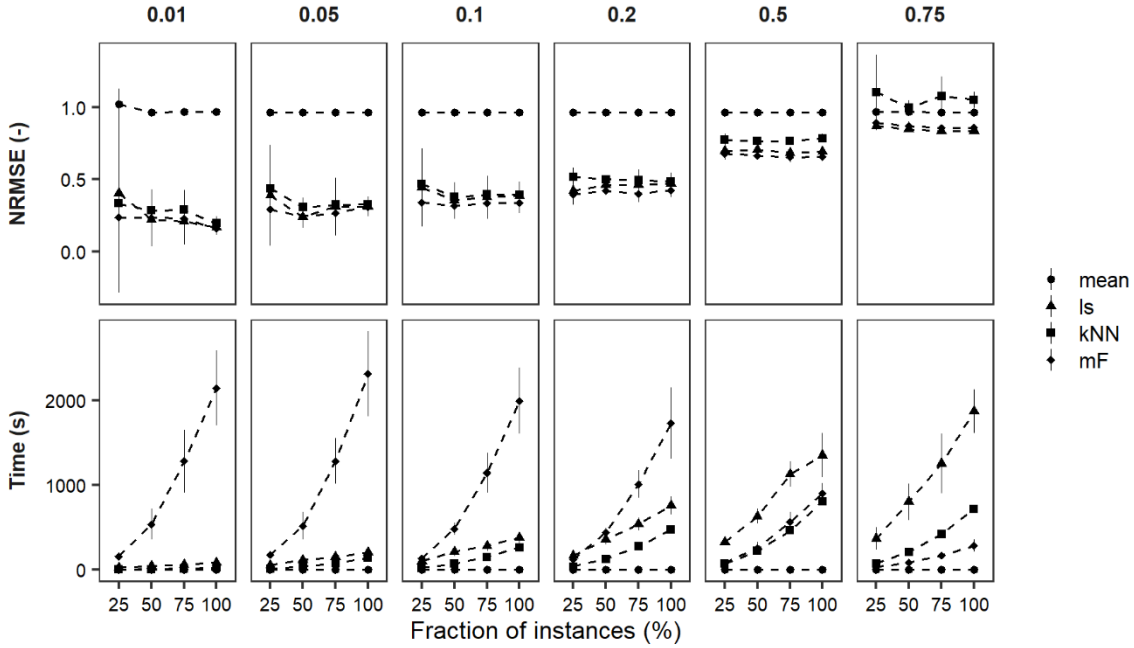

Figure 5.3: Effect of number of instances for a fixed number of variables. *More instances have a limited impact on performance, but do affect computation time. The top row (NRMSE) represents performance, while the second row represents the required computation time. Columns represent the different degrees of missing data used. Data of 10 repetitions (variable number of instances for 10 variables, different missing values) are reported. Symbols represent the average for each combination, while vertical lines represent the standard deviation. Methods:* mean*: mean imputation;* mF*: the missForest algorithm;* kNN*: k nearest neighbours and* ls*: iterative least squares. NRMSE: Normalised Root Mean Squared Error.*

5.3.3 Dimensionality variability

Inclusion of additional variables was expected to increase imputation performance, despite the underlying reduction in sample size and a potential to increase model overfitting. The latter is consequential to the consideration of a high number of variables to explain or describe the patterns within the data and is characterised by a reduced accuracy outside its training range. Hence, despite an increased explanatory power by including additional variables, a decrease in imputation accuracy can be obtained. Still, dimensionality clearly affected imputation performance, with a general decrease in error following an increase in dimensionality (Figure 5.4). Only *kNN* did not show a monotonous increase in performance when 50 % or more of the data was missing, but rather performed worst at intermediate dimensionality (0.97 ± 0.06 at N_{var} = 5 versus 1.05 ± 0.06 at N_{var} = 10, with 75 % missing).

The saturated model indicated that a significant overall interaction existed and that inclusion of a main effect and interaction with imputation method significantly improved model fit (p < 0.001). Interaction coefficients indicated a higher effect of dimensionality on *mF* ($\beta_{mF:Var}$ = -0.067) compared to *ls* and *kNN* ($\beta_{ls:Var}$ = -0.044 and $\beta_{kNN:Var}$ = -0.025), causing the overall significant differences between *ls*, *kNN* and *mF* in the baseline performance (see earlier section) to disappear. Still, they provided significantly higher performance than *mean*, regardless of missing data and dimensionality (all p < 0.001), except for *kNN* at 75 % missing values and only 5 variables. In contrast, with 50 % or less of the data missing and only 5 variables, *kNN* performed similarly as *ls* and *mF*, yet performance discrepancy increased when 10 (*mF*) or 15 (*ls*) variables were available (p < 0.05). Differences between *ls* and *mF* were mostly non-significant, except at increased dimensionality (≥ 10 variables) and elevated degrees of missing data (≥ 20 %).

Both *mF* and *kNN* showed maximal required computation time at intermediate dimensionality (N_{var} = 10, up to 2100 ± 400 s for *mF*) and minimal at increased dimensionality (N_{var} = 15, 220 ± 60 s for *mF* at 1 % missing) (Figure 5.4). Surprisingly, the latter did not result in a clear change in computation time for *kNN* or *mF* along the range of missing data, while reduced dimensionality showed a similar pattern as observed in Figure 5.2. In contrast, both *mean* and *ls* were not clearly affected by dimensionality nor the degree of missing data (Figure 5.4).

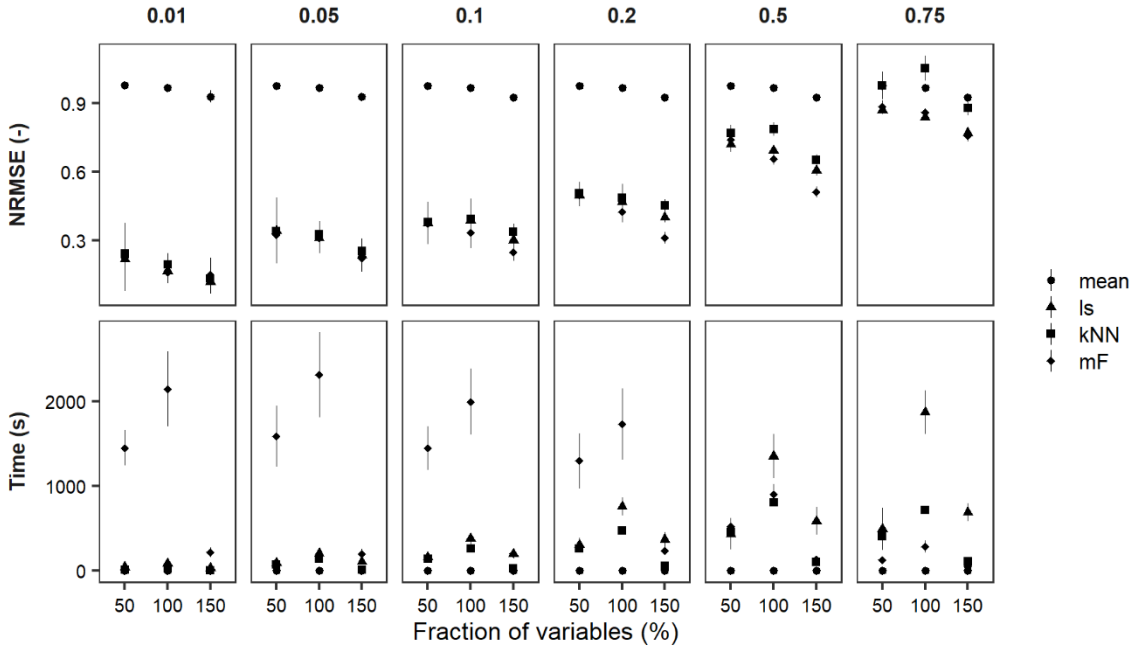

Figure 5.4: Effect of dimensionality on performance and required time of four imputation methods. *Not only the number of variables, but also sample size is different among data sets. The top row (NRMSE) represents performance, while the second row represents the required computation time. Columns represent the different degrees of missing data used. Data of 10 repetitions (maximum number of instances for a specific number of variables, different missing values) are reported. Symbols represent the average for each combination, while vertical lines represent the standard deviation. Methods:* mean*: mean imputation;* mF*: the missForest algorithm;* kNN*: k nearest neighbours and* ls*: iterative least squares. NRMSE: Normalised Root Mean Squared Error.*

5.3.4 Optimisation

Preliminary assessment showed that additional typological information did not result in improved imputation performance (see Appendix, Figure B.1), hence this was not considered for further elaboration. In contrast, altering hyperparameter settings often improved performance and was therefore included in subsequent analyses. Specific effects of each individual hyperparameter were considered being beyond the current scope and merit additional study.

By default, *kNN* considers five neighbours, yet the optimised k_{nn} value ranged from 1 up to 47, with a median value of 9. Almost 33 % of the k_{nn} values were equal to or lower than 5, while another 33 % ranged from 15 up to 47. In general, data sets with low rates of missing data ($f_{MD} \leq 10$ %) supported improved imputation when low k_{nn} values were applied ($k_{nn} \leq 10$) and vice versa for data sets with elevated rates of missing data (Figure 5.5).

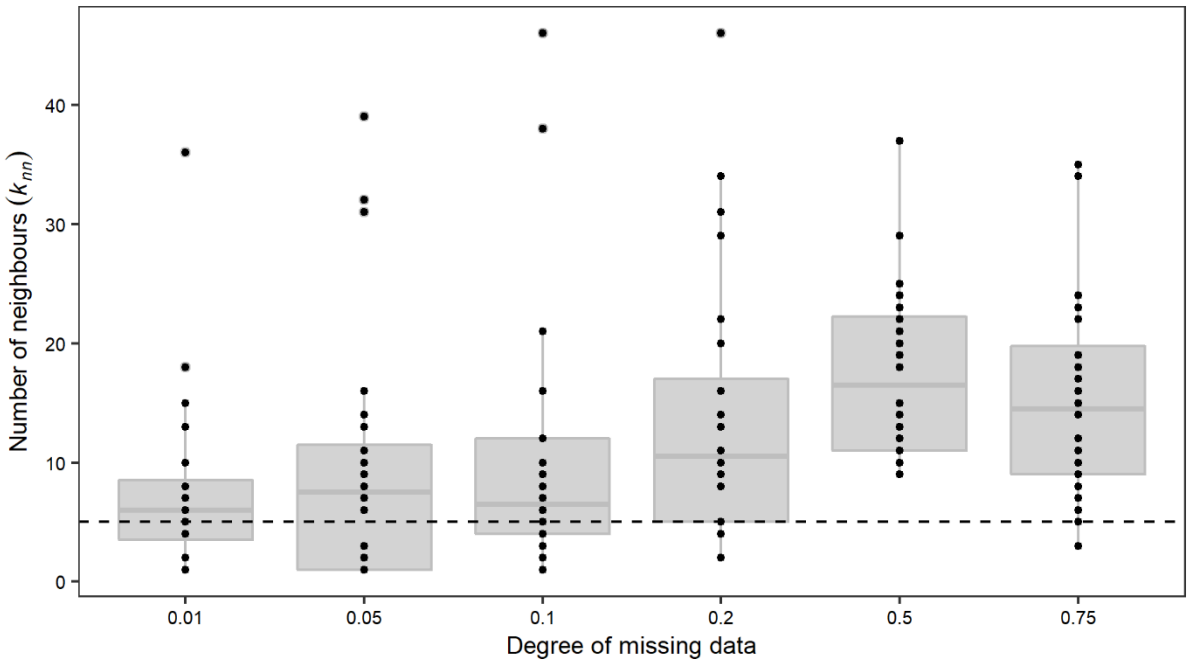

Figure 5.5: Selected number of neighbours to be considered after optimisation. *Optimised* k_{nn} *values were determined for the six classes of missing data (0.01, 0.05, 0.1, 0.2, 0.5 and 0.75), represented by two repetitions of each possible combination of sample size (number of instances) and dimensionality (number of variables), hence a total of 144 data sets. Values range from 1 up to 47, with low values being selected when imputing data sets with limited rates of missing values and vice versa for data sets with high rates of missing values. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range.*

Similar patterns could not be identified for *mF* due to the simultaneous alteration of three hyperparameters during the iteration process, yet observations suggested that the majority of the data sets was imputed with higher accuracy when less individual trees were constructed and more variables were randomly selected at each split (see Table 5.2 and Appendix, Figure B.5 and Figure B.7). For instance, *ntree* ranged from 5 up to 225, with the majority of the data sets requiring less than the default number of trees (i.e. *ntree* = 100).

Indeed, 75 % of the data sets required 84 trees or less to improve imputation accuracy, while median values for *mtry* were similar (N_{var} = 5) or higher (N_{var} > 5) than the default value (Table 5.2). Quantitative improvements in NRMSE values were, in general, smaller than 0.25 and relatively unaffected by the rate of missing data and the default performance for both methods (Figure 5.6). On average, the absolute decrease in NRMSE values between the default and optimised imputation settings was 0.05 ± 0.05 (N = 144) for both methods.

***Table 5.2: Summarising statistics for the optimised hyperparameter values of* mF.** *Optimised values were determined for the six classes of missing data (0.01, 0.05, 0.1, 0.2, 0.5 and 0.75), represented by two repetitions of each possible combination of sample size (number of instances) and dimensionality (number of variables), hence a total of 144 data sets. In general, the majority of the data sets benefit when imputation is performed with less individual trees (*ntree*) and more variables to be considered for each split (*mtry*).*

	Default	Min	Q1	Median	Mean	Q3	Max
ntree	100	5	25	50	62	84	225
mtry (N_{var} = *5*)	2	1	1	2	2	3	4
mtry (N_{var} = *10*)	3	1	2	3	4	5	9
mtry (N_{var} = *15*)	3	2	4	6	7	9	14
nodesize	1	1	1	2	2	2	6

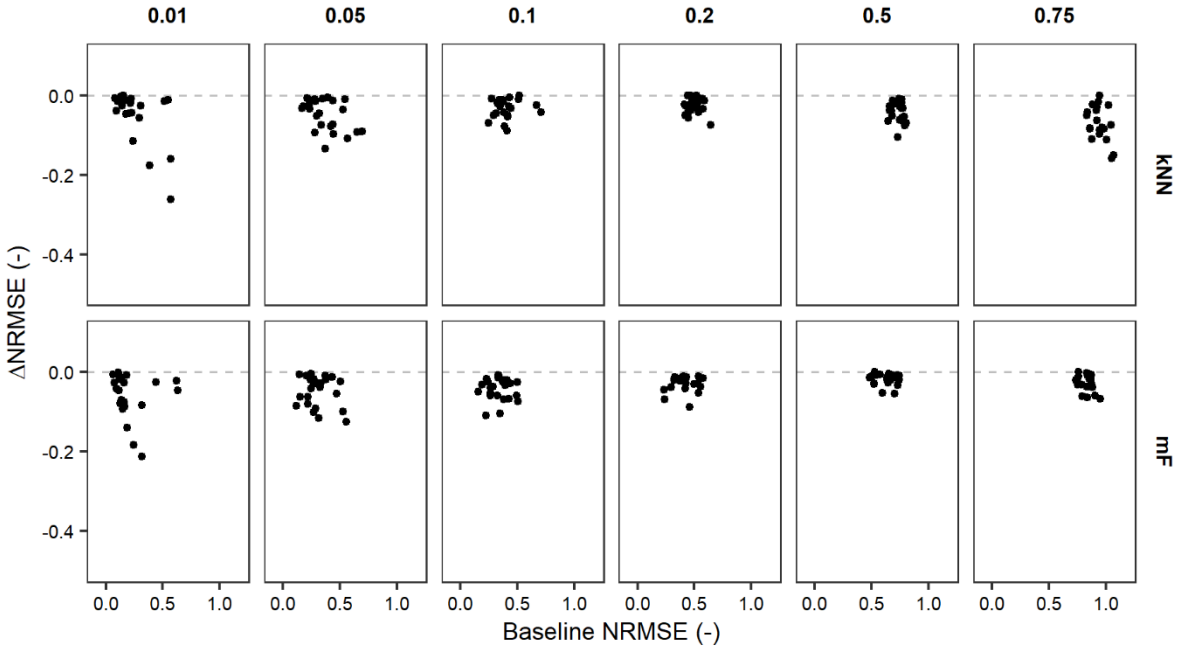

Figure 5.6: Error reduction following optimisation of hyperparameter settings of the* kNN *and* missForest *algorithm. *The difference is calculated as the NRMSE value after optimisation minus the NRMSE value in case of default settings (Baseline NRMSE). The horizontally grey dotted line represents the reference condition (i.e. no change in NRMSE), with symbols below it reflecting an improvement of performance and symbols on it reflecting a steady state. Selected hyperparameters included the number of neighbours (*kNN*) and the number of individual trees, number of variables to be considered for each split and final nodesize for missForest (*mF*).*

5.4 Discussion

5.4.1 Performance evaluation

The high degree of stability obtained by imputing the mean value (*mean*) illustrates its reliability as it is hardly affected by the degree of missing data, the number of instances, nor the number of variables. Along with its simplicity and low computation times, *mean* represents a pragmatic imputation method, though performed worst in this comparative study despite outperforming other methods in literature (Shrive *et al.*, 2006). When facing high degrees of missing values (f_{MD} > 75 %), *mean* appears to become a valid approach, potentially due to the lack of sufficient information for *ls*, *kNN* and *mF*. Yet, imputing high rates of missing values greatly affects the estimation of population statistics and associations (García-Laencina *et al.*, 2010; Little and Rubin, 2002; Penone *et al.*, 2014), which increases the chance of imputing values that deviate strongly from the actual value, as illustrated by the increased error for *ls*, *kNN* and *mF*. Still, mean imputation narrows the variable's distribution and results in an underestimation of the standard deviation and the population's variance, thereby additionally affecting subsequent analyses like PCA and habitat suitability model development (Brock *et al.*, 2008; Liew *et al.*, 2011).

Narrowing causes more distant values to become underrepresented and, hence, potentially ignored during model development, inhibiting both the interpretation of descriptive models and the extrapolation of predictive models. Therefore, some authors support the idea of considering data imputation and model performance at once, as higher imputation accuracy does not necessarily warrant improved model performance (Brock *et al.*, 2008; García-Laencina *et al.*, 2010; Luengo *et al.*, 2012). However, this should be done with care as it might favour conservative imputation approaches, thereby artificially inflating performance metrics.

Along with *mean*, *ls* is not subject to hyperparameter-tuning and is only limitedly affected by the number of iterations to be performed. Despite the iterative approach, *ls* provides visually similarly performance as *mF* and *kNN* at low degrees of missing data (f_{MD} ≤ 20 %), while outperforming *kNN* at higher degrees of missing data (f_{MD} ≥ 50 %). However, in spite of the global approach of *ls* (Bø *et al.*, 2004), computation time remains tends to increase greatly along the degree of missing values. Hence, a multivariate regression approach provides a promising perspective for imputing multidimensional environmental data, especially when extension beyond linear associations is possible (e.g. GLM-based).

Being outperformed by *ls* and *mF* at high levels of missing values classifies *kNN* as an intermediately performing method, thereby complying with literature (Celton *et al.*, 2010; Schmitt *et al.*, 2015). Moreover, at low degrees of missing data, *mF* tends to provide significantly better performance than *kNN*, indicating that, under default hyperparameter settings, *mF* provides overall better performance.

The power of *mF* resides in the combination of several individual trees (i.e. a bagged imputation technique) and an iterative approach that allows to update the imputed values (Waljee *et al.*, 2013), hence explaining the high computation times required for *mF*. Clearly, *mF* requires more time than *mean*, *ls* and *kNN*, except at high levels of missing data (f_{MD} > 50 %), due to combining global and local associations. The observed reduction in computation time as more data became missing is a potential consequence of reduced dimensional space, providing a basis for a trade-off analysis between required data dimensionality and computational time. Contrasting this decrease, *kNN* shows an increase in computation time, which is a potential consequence of requiring a more intensive search for imputing all missing data points and finding the appropriate neighbours.

In short, *mF*, *kNN* and *ls* provide relatively low overall imputation errors at low levels of missing data (even without optimisation of *mF* and *kNN*), demonstrating that a single-best approach does not exist (Brock *et al.*, 2008; Celton *et al.*, 2010; Liew *et al.*, 2011). For instance, *mF* provides overall relatively high accuracies, yet when computation time is restricted, *ls* represents a valid alternative at low rates of missing values.

5.4.2 Sample size and dimensionality

Alterations in sample size and dimensionality provided the expected pattern of reduced performance following a decrease in either sample size or dimensionality. Indeed, negative coefficients of the main effects were obtained (see Appendix, Table B.6 and Table B.7), reflecting a general decrease in error when sample size and/or dimensionality is increased. Effects differed among the imputation methods, but were generally stronger for *kNN*. Consequently, these observations suggest that removal of instances or variables prior to data imputation is only to be considered when additionally providing a reduction in the fraction of missing data. Similarly, inclusion of additional instances and/or variables is only beneficial when the degree of missing values does not increase, as this counteracts the positive effect of augmented sample size and dimensionality. Moreover, depending on the type of data included, error reduction might be relatively limited. For instance, introduction of typological data had a minor effect on imputation performance and even caused higher errors to occur (see Appendix, Figure B.1). Yet, when high errors are expected (e.g. at high levels of missing data), additional data can support slightly better performance.

5.4.3 Fine-tuning via optimisation

In contrast to *mean* and *ls*, both *mF* and *kNN* make use of hyperparameters to support data imputation. Performance of *mF* is affected by various hyperparameters, including number of trees (*ntree*), number of variables for each split (*mtry*) and the nodesize to be considered, while *kNN* is only affected by k_{nn}, reflecting the number of neighbours for calculating the weighted average. Results showed that optimisation is highly case-specific (see also Appendix B.3) as hyperparameter settings and related performance rely on the intrinsic correlations within the data (Brock *et al.*, 2008). Consequently, no specific set of hyperparameter settings can be specified, yet some general guidelines for alternative settings can be inferred:

1. The rate of missing data affects the optimal number of neighbours of *kNN*. Low rates (up to 10 %) will perform well with the default value of k_{nn} = 5 and a search range of ± 5. Intermediate rates (20 %) can be centred around k_{nn} = 10 with a range of ± 10. Lastly, high rates (50 % and up) cover a wide range of potential optimal values, yet a starting point could be k_{nn} = 15 with a range of ± 10.
2. The number of individual trees can be slightly reduced, with a positive impact on computation time. For instance, *ntree* = 80 can be considered as starting point, decreasing computation time by 20 %, due to its linear relationship with *ntree* (Stekhoven and Bühlmann, 2012).
3. The number of variables to be considered for each split can be increased. For instance, the square root of the original number of variables can be replaced by division by 2.
4. Nodesize is relatively irrelevant when aiming to obtain improved accuracy. It might, however, reduce complexity and increase transparency of individual trees and should only be altered if interpretability is an additional goal.

Nevertheless, performance can be improved for both *mF* and *kNN* (Δ_{NRMSE} = -0.05 ± 0.05), represented by a maximum absolute difference in NRMSE up to 0.35 and 0.34, respectively. These improvements are similar regardless the degree of missing data nor the applied method, suggesting that the original difference in performance remains present with overall best imputation accuracy provided by *mF*. Still, despite the increased performance, methods without an optimisation-option or already including optimisation might be favoured over *mF* and *kNN*, solely because of the additional increase in computation time of the latter (Brock *et al.*, 2008).

5.4.4 Implications for field-based research

A potentially interesting consequence of these observations represents the possibility to allow incomplete data to be present within the assessment data set, supporting the collection of more instances and/or variables. Likewise, data collection campaigns can be designed to randomly select data points that can be excluded during sampling as a way to save both time and money. For instance, assuming that the collection of each data point is equally expensive and time consuming, increasing the number of instances from 8 000 with 5 % missing data to 12 000 with 10 % missing values, allows that within the collected 4 000 instances 20 % of the values are missing, representing about 800 data points. Collection of information on all data points (hence no missing data) within the same time and budget, would limit the amount of instances to be collected to 3 200. Hence, a proper design allows for more information to be collected by allowing a certain degree of missing values, preferably assigned randomly in advance. Including variable-specific information related to costs and timing allows for testing multiple random missingness schemes in order to optimise the time-budget-information nexus. Yet, one should always be aware that data imputation does not legitimately equals proper data collection and that each imputation causes a distortion of the hidden patterns (Nogueira *et al.*, 2007). Hence, results obtained through data imputation should be interpreted with care, as these distortions can range from being relatively small (e.g. minor changes in variable correlations with an overall low NRMSE) up to being disruptive (e.g. decreasing variable range with 50 %). Yet, the performed case studies suggested that only imputation of the mean created distinct changes in variable distributions, although the extent of most variables might have masked smaller distortions (see Appendix B.4).

Nevertheless, the complete absence of missing values in publicly available data is hard (if not impossible) to obtain as the amount of data continues to increase along with the pressure to make data publicly available (Gibert *et al.*, 2018a). Yet combining data from different research questions unavoidably leads to missing values as a consequence of not-recording. Moreover, even though continuous monitoring is becoming less budget-intensive, it is often affected by (i) low temporal resolution and (ii) defects, which create gaps within time series that limit the capture of variable dynamics and frequency distributions (Giustarini *et al.*, 2016). In contrast, funds for specific environmental monitoring campaigns are decreasing globally and highlight a need for (i) cheaper monitoring technology and (ii) well-structured data sets with appropriate commentary (Sprague *et al.*, 2017). This illustrates the need within the water management sector for imputation techniques to avoid both information and investment loss.

5.4.5 Contribution to the study objective

The aim of this chapter was to elucidate the differences between a selection of available imputation techniques in order to tackle the relatively high degree of missing data in the physicochemical data enclosed in the Limnodata Neerlandica. Throughout the chapter, a collection of complete data sets were derived from the original database and exposed to artificial random data point removal in order to infer technique-specific imputation errors. Moreover, by considering a variety of potential data set dimensions, a more pronounced basis was created to bring forward a specific imputation technique for further data cleaning within the overall study objective (see Section 1.2.1). It should remain clear that this chapter contributes mostly to the overall study objective, while providing suggestions for application outside the considered framework. More specifically, it is recommended to perform similar analyses with different combinations of environmental variables to support empirical technique selection.

The chapter provides a solution for the high degree of missing data (93.7 %, see Section 4.2.1.3) that occurs within the combined physicochemical and macrophyte occurrence data. As this was mostly caused by variables with hardly any information (i.e. only 6 variables contained information for more than 50 % of all instances), a reduction in the number of variables positively affected the overall degree of missing data. Yet, variable reduction aiming to obtain only complete cases caused an unwanted reduction in the dimensionality of the observed environmental domain. Hence, the imputation of missing data based on available data provided an alternative solution.

The selection of imputation techniques was limited to the methodologies that provided single-value imputation, i.e. providing a single complete data set after replacing the missing data points. More advanced multiple-value approaches exist, though these often require the individual analysis of each new data set (Faris *et al.*, 2002; Lee and Carlin, 2010; Schmitt *et al.*, 2015). As the main study aim entailed the development of several species-specific models, such multiple imputation would increase the computation and analysis time tremendously. Therefore, a selection of single-value techniques was made based on literature and technique-specific characteristics.

In general, technique application supported the expectations at the level of (i) data set characteristics and (ii) technique-specific characteristics. For instance, increased data dimensionality and sample size positively affected imputation accuracy, while lowest imputation errors were mostly obtained by random forests. Moreover, the latter provided better performance than mean imputation for f_{MD} values up to 50 %. Therefore, the *missForest* technique is considered for subsequent imputations, while aiming to reduce the degree of missing data to 50 % (being below the 90 % reported by Madley-Dowd *et al.* (2019)).

5.5 Conclusion

Four imputation methods with different degrees of application complexity were selected, providing a mix of transparent and so-called black-box methods while simultaneously representing well-known and more recent methods to impute environmental data. This selection is far from exhaustive, but provides a sound addition to the data pre-processing options when dealing with environmental data. The results showed that the random-forest based *missForest* algorithm outperforms other methods, while the regression-based *least squares* and similarity-based *k nearest neighbours* approaches provide valid alternatives when computation time is restricted and less than 20 % of the data is missing. Moreover, imputation accuracy improves when (1) more variables are included rather than adding instances and (2) an iterative procedure of hyperparameter optimisation is conducted. It has to be noted, however, that the comparative nature of this study is limited by the fact that both temporal and logical data were not included, aside from the assumption that the missing data mechanism reflects a missing completely at random (MCAR) design, yet similar results are to be expected for missing at random (MAR). Despite these limitations, valuable observations across different conditions (sample size and dimensionality) were obtained, supporting future data pre-processing within the field of environmental data analysis and habitat suitability model development.

6

Speed-performance trade-off in threshold selection during data pre-processing[4]

Highlights

- Eliminating outliers and redundant variables decreased model performance
- Avoiding false absences improved model performance
- Data removal supported faster model development
- Combinatory data pre-processing increased performance and computation time

[4] This chapter is based on Van Echelpoel, W.; Bruneel, S. and Goethals, P. L. M. (in preparation) Speed-performance trade-off in threshold selection during data pre-processing

Abstract

Real-world data requires cleaning prior to performing in-depth analyses and concluding on qualitative results. During data cleaning, associations among variables are analysed, the reliability of recorded values is registered and irrelevant or erroneous data are removed. This positively affects the quality of the training data, despite requesting tremendous temporal and budgetary investments, by improving the discoverability of patterns within it, thereby supporting the development of accurate and simple models. Progress in the field of data mining increases rapidly, yet mainly focuses on specific and novel data mining techniques rather than optimising data preparation, causing an artificial mismatch between the supplied low-quality data and the demanded high-quality data. Here, four different data pre-processing options are introduced and discussed. Outliers, false absences and variables that are correlated or irrelevant are identified and excluded from the training data to infer the effect of data pre-processing on conditional random forest performance and required computation time. Each method is characterised by a user-defined threshold, causing results and conclusions to be highly case-dependent. A visual trade-off analysis of model performance, required computation time and data set characteristics supported the identification of thresholds for the elimination of outliers (τ_o = 3), false absences (τ_a = 5 %), correlated variables (τ_c = 0.7) and irrelevant variables (τ_i = 10 %). Serial combinatory data pre-processing improved model performance with net AUC increases up to 0.1, though simultaneously caused a drastic increase in computation time. Nevertheless, final model performance ranged up to AUC values equal to 0.85 and increased even more when the external test data was devoid of false absences. These results indicate that overall data pre-processing positively affects model performance at the expense of computation time and that niche-based exclusion of false absences is crucial to comply to the equilibrium assumption within correlative habitat suitability modelling. Moreover, they illustrate that the abovementioned thresholds can be used in future studies, while highlighting that inclusion of the implemented threshold within scientific reports is essential to improve replicability.

6.1 Setting the scene

Chapter 5 already illustrated how missing values within environmental data sets could be tackled. Yet, additional actions are needed to perform proper data cleaning prior to deriving qualitative results (Gueta and Carmel, 2016; Zhang *et al.*, 2003). Data cleaning positively affects the quality of the training data by improving the discoverability of patterns within it, thereby supporting the development of accurate and simple models (Kotsiantis *et al.*, 2006; Maldonado *et al.*, 2015). Progress in the field of environmental data mining has been increasing rapidly, with a main focus on the development of specific and novel techniques (Zhang *et al.*, 2003). The resulting delayed interest in the value of qualitative data has steered the improved awareness on data importance and has increased the application and development of data pre-processing methods. Unfortunately, comparative studies and detailed analyses of the effect of data pre-processing thresholds on data availability and model performance remain rare.

On the one hand, noise introduced by outliers distorts the factual representation of environmental ranges caused by artificial range extension. The nature of these outliers ranges from natural variability to erroneous notation and can lead to reduced model accuracy. More specifically, outliers related to reported species presence create a basis to overestimate (1) the species' realised niche and (2) the potential geographical distribution (Lobo *et al.*, 2010; VanDerWal *et al.*, 2009). Implementation of outlier identification varies among studies due to a lack of guidelines and comparative research. For instance, Gobeyn *et al.* (2017) applied visual inspection of box plots, histograms and dot plots to identify outliers in a subjective manner, while VanDerWal *et al.* (2009) considered a range of environmental extents to determine the best-performing one.

Opposite of eliminating outliers stands the identification of ambiguous information among highly similar instances. For example, false absences caused by non-detection of a rare species or non-occupation of a suitable habitat due to dispersal limitation insinuate an unsuitable habitat (Anderson and Raza, 2010). Similarly, false presences due to misidentification or a lagged response to altered conditions have the potential to untruly extend the species' realised niche (Lobo *et al.*, 2010). Generally, efforts to avoid the inclusion of false absences and presences is biased towards the former as most studies rely on the assumption that the error among recorded presences is negligible (up to non-existing). False absence rates are expected to be higher than false presence rates due to a complex interplay of biotic interactions, historic events, dispersal limitations and dynamic physiological processes, making it hard to confirm true absences (Lobo *et al.*, 2010). Consequently, most occurrence-based species distribution studies make use of pseudo-absences rather than true absences to contrast confirmed presences (Chefaoui and Lobo, 2008; Phillips *et al.*, 2009). These pseudo-absences entail all locations where species have not been observed, thereby combining both true and false absences.

On the other hand, irrelevant and correlated variables have limited value in correlative model development as they increase data dimensionality, required computation time and model complexity (Kotsiantis *et al.*, 2006). Identification of relevant variables relies on expert knowledge or on preliminary correlative model(s) and subsequent assessment of variable importance. Reduction of data dimensionality and model complexity by eliminating irrelevant variables is claimed not to significantly affect model accuracy. For instance, Fox *et al.* (2017) studied the effect of score-based variable selection on model performance and observed that in the case of random forests, no significant change in performance was noted. This illustrates that variable selection mostly aims at complexity reduction (i.e. model regularisation) rather than improving accuracy.

Analogously, correlated variables represent similar information and indicate that model complexity can be reduced by selecting either one. Often, this selection is based on ecological knowledge, relation with the response variable or even variable importance. For instance, Forio *et al.* (2018) considered the degree of missing data as basis for correlated variable removal, while Sauer *et al.* (2011) relied on expert knowledge to determine which variable to retain. Within occurrence-based species distribution studies, frequently applied correlation threshold values for input variable selection vary between 0.7 (e.g. Gobeyn *et al.* (2017), Van Echelpoel and Goethals (2018)) and 0.8 (e.g. Forio *et al.* (2018), Sauer *et al.* (2011)), though often no strict threshold is reported.

A common characteristic among these pre-processing techniques, is the inclusion of one (or more) technique-specific threshold(s). These thresholds need to be defined by the user prior to technique implementation, while affecting the final result. Still, despite the widespread application of data pre-processing in ecological research, effects of data cleaning, threshold value selection and combinatory data pre-processing on both model performance and computation time remain relatively understudied (Gueta and Carmel, 2016). Moreover, threshold values are only limitedly reported and often case-specific, underlining the need for a solid conceptual framework to support decision-making (Kotsiantis *et al.*, 2006; Zhang *et al.*, 2003).

Within this chapter, attention is given to four data pre-processing techniques to select instances or variables. The aim is to assess the effects of technique-specific threshold selection on model performance and the required computation time and to suggest a single threshold for future combinatory data pre-processing. More specifically, this chapter complies to objective 2.2 as defined in Chapter 1 and completes the answer to RQ2.1. Hence, this chapter concludes with a statement on the suggested technique-specific threshold values to be used for data quality improvement and future model development. These values are not claimed to be the holy grail for all future environmental data science projects. Rather, this study provides an illustration of how threshold selection can be performed.

6.2 Materials and method

6.2.1 Characterisation of the data

Data within the Limnodata Neerlandica was subsampled to contain spatiotemporally referenced observations of macrophytes and the prevailing physicochemical conditions (see Section 4.2.3.2), providing information on 4344 instances, 174 variables and 576 macrophytes (Figure 4.3). Physicochemical data was characterised by a high number of variables that contained limited information, causing a high degree of missing values (93.7 %) and therefore requiring further reduction. The degree of missing data was reduced to 49.7 % (with 50 % being considered manageable for imputation, see Chapter 5) by stepwise removal of the variable or instance that contributed most to the overall reduction, providing information on 4158 instances and 20 variables (see Appendix, Figure C.1 and Table C.1). Subsequently, missing data was imputed by using the *missForest* algorithm with default settings (see Chapter 5).

For each instance, a presence/absence statement reflecting macrophyte occurrence was available, yet overall prevalence was often below 2.4 % (i.e. 100 instances in total). These low-prevalence macrophytes were left out, while remaining macrophytes were double-checked for representing plants with a main aquatic life-stage. This resulted in a final data set of only 58 different macrophyte species, along a prevalence range between 2.4 and 41 %. Analyses were performed for all macrophytes, yet for brevity reasons a subset of five macrophytes was selected, covering (1) the observed prevalence range (2.4-41 %), (2) different growth forms (emergent, submerged and floating) and (3) origin (native, alien), being presented in Table 6.1 (and Appendix, Table C.2). Data preparation and all subsequent calculations and modelling activities were performed in RStudio (R Core Team, 2016; RStudio Team, 2015), while making use of the packages *missForest*, *party* and *PresenceAbsence* (Freeman and Moisen, 2008a; Stekhoven, 2013; Strobl *et al.*, 2009a).

Table 6.1: Characterisation of the macrophyte subset. *Five macrophytes were selected to cover the observed prevalence range, different growth form and origin. Note that origin here is considered for western Europe in general and that classification into* native *or* alien *is highly dependent on the considered timeframe.*

Macrophyte	Prevalence (%)	Growth form	Origin
Phragmites australis	41	Emergent	Native
Lemna minor	27	Floating	Native
Ceratophyllum demersum	18	Submerged	Native
Mentha aquatica	11	Emergent	Native
Lemna minuta	3	Floating	Alien

6.2.2 Preliminary assessment

Based on the abovementioned data set, a preliminary study was implemented to determine the minimum number of trees (*ntree*) to be developed within the conditional random forest (CRF) as well as the number of repetitions to be carried out (*n_rep*). First, *ntree* was defined to range between 50 and 1000 (step size equal to 50) to infer the stabilisation point of the developed forest. Secondly, the influence of repetitions on variance reduction was examined up to 30 repetitions, aiming to define the number of required repetitions for the AUC stabilisation. For each parameter, visual assessment was performed to infer the stabilisation point and, hence, which values to use for subsequent analyses. Due to the this specific construction, a total of $k_{cv} \cdot n\_rep$ individual AUC scores was obtained (with k_{cv} = 5 representing the cross-validation) and combined into an overall AUC score.

6.2.3 Data pre-processing techniques

With the settings inferred from the preliminary assessment, CRFs were developed, which involved the testing of the effect of further data pre-processing on model performance. Due to the specific construction of the random forest algorithm, it is expected that the inclusion of both outliers and correlated variables has a limited effect (Breiman, 2001; Fox *et al.*, 2017; Vezza *et al.*, 2015). However, the reduction of incorrect and irrelevant information improves model regularisation by reducing model complexity and therefore relies on the trade-off between data and model complexity. Consequently, model regularisation encompasses appropriate instance and variable selection (i.e. identifying and eliminating outliers, false absences, correlated and irrelevant variables).

6.2.3.1 Selection of instances

Detection of outliers

Practical implementation of outlier identification and removal starts with considering the original $N_{inst} \times N_{var}$ dataset ($\boldsymbol{D}$) and creating a new, equally-dimensioned matrix $\boldsymbol{O}$. For each variable X_j ($j \leq N_{var}$) the first and third quartile are defined ($Q_{j,1}$ and $Q_{j,3}$, respectively) as well as a user-specified range threshold ($\tau_{o,j}$). Subsequently, Equation 6.1 is applied to $d_{i,j}$ ($\in \boldsymbol{D}$) and an outlier dummy score (1 if considered outlier, 0 if not) is assigned to $o_{i,j}$ ($\in \boldsymbol{O}$). Finally, outlier dummy scores are summed for each instance, causing instances that exceed the pre-specified threshold α_o (i.e. $\sum_{j=1}^{N_{var}} o_{i,j} \geq \alpha_o$) to be removed from the data set.

To assess the effect of range selection, $\tau_{o,j}$ was set to range from 0 (high degree of removal) up to 15 (low degree of removal) with a step size equal to 1, without being variable-specific (i.e. $\tau_{o,j} = \tau_o$). Meanwhile, α_o was fixed to 1, reflecting the idea that an instance with an outlier score in 1 variable becomes less reliable and should, therefore, be removed from the data.

$$o_{i,j} = \begin{cases} 1, & d_{i,j} < Q_{j,1} - \tau_{o,j} \cdot (Q_{j,3} - Q_{j,1}) \\ 0, & Q_{j,1} - \tau_{o,j} \cdot (Q_{j,3} - Q_{j,1}) \leq d_{i,j} \leq Q_{j,3} + \tau_{o,j} \cdot (Q_{j,3} - Q_{j,1}) \\ 1, & d_{i,j} > Q_{j,3} + \tau_{o,j} \cdot (Q_{j,3} - Q_{j,1}) \end{cases} \quad \text{(Equation 6.1)}$$

With $d_{i,j}$ the value of the j-th variable of the i-th instance, $o_{i,j}$ the outlier dummy score of $d_{i,j}$, $Q_{j,1}$ the first quartile of the j-th variable, $Q_{j,3}$ the third quartile of the j-th variable and $\tau_{o,j}$ the user-specified threshold for the j-th variable.

Detection of pseudo-absences

Instance selection based on false absence identification started with the separation of presences and absences. Based on the absence data set $\boldsymbol{D_{abs}}$ ($N_{abs} \times N_{var}$) a new, equally-dimensioned matrix $\boldsymbol{A}$ is created. For each variable X_j ($j \leq N_{var}$) distribution percentiles ($P_{\frac{\tau_{a,j}}{2}}$ and $P_{\left(1-\frac{\tau_{a,j}}{2}\right)}$) of the occupied environmental domain (i.e. presence data set, $\boldsymbol{D_{pres}}$) are defined, including a user-specified range threshold ($\tau_{a,j}$). Subsequently, Equation 6.2 is applied to $d_{i,j}$ ($\in \boldsymbol{D_{abs}}$) and an absence dummy score (1 if considered potential true absence, 0 if not) is assigned to $a_{i,j}$ ($\in \boldsymbol{A}$). Finally, absence dummy scores are summed for each instance, causing instances that exceed the pre-specified threshold α_a (i.e. $\sum_{j=1}^{N_{var}} a_{i,j} \geq \alpha_a$) to be maintained in the absence data set. This approach is visualised in Figure 6.1 for two variables, but can be easily extended to higher dimensions. Lastly, the presence and updated absence data are merged into a single data set for model training.

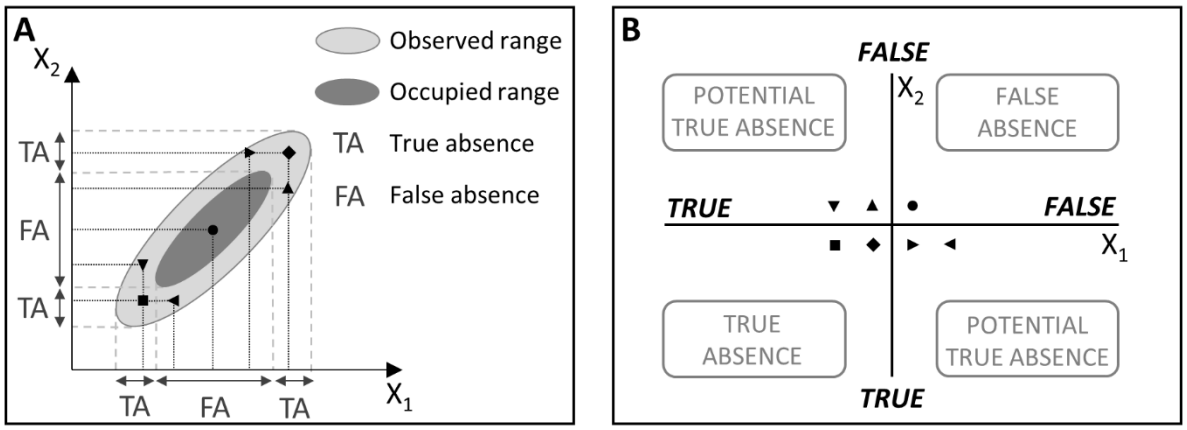

Figure 6.1: Illustration of the false absence concept. *Each absence can be classified as a true absence, a potentially true absence or a false absence when related to occupied environmental niche. A: Situation of observed absences along two environmental gradients (X_1 and X_2) with respect to the observed environmental domain (light grey) and occupied environmental domain (dark grey). B: Classification of the observed absences from (A) based on being true or false in the individual environmental gradients. The value of τ_a determines the extent of the occupied range in (A), while the value of α_a influences which potential true absences from (B) are ultimately included in the model training data.*

To assess the effect of range selection, $\tau_{a,j}$ was set to range from 0 (high degree of removal) up to 0.15 (low degree of removal) with a step size equal to 0.01, without being variable-specific (i.e. $\tau_{a,j} = \tau_a$). Meanwhile, α_a was fixed to 1, reflecting the idea that an instance with an absence score in 1 variable is situated outside the realised environmental niche and should be kept in the data set. Hence, potential true absences are included in the resulting absence data set (see Figure 6.1) and subsequently used as training data.

$$a_{i,j} = \begin{cases} 1, & d_{i,j} < P_{\frac{\tau_{a,j}}{2}} \\ 0, & P_{\frac{\tau_{a,j}}{2}} \leq d_{i,j} \leq P_{\left(1-\frac{\tau_{a,j}}{2}\right)} \\ 1, & d_{i,j} > P_{\left(1-\frac{\tau_{a,j}}{2}\right)} \end{cases} \qquad \text{(Equation 6.2)}$$

With $d_{i,j}$ the value of the j-th variable of the i-th instance, $a_{i,j}$ the absence dummy score of $d_{i,j}$, $P_{\frac{\tau_{a,j}}{2}}$ the lower percentile of the j-th variable, $P_{\left(1-\frac{\tau_{a,j}}{2}\right)}$ the upper percentile of the j-the variable and $\tau_{a,j}$ the user-specified threshold for the j-th variable.

6.2.3.2 Selection of variables

Identification of correlated variables

Correlation-based dimensionality reduction starts by considering the original $N_{inst} \times N_{var}$ dataset ($\boldsymbol{D}$) and the construction of a $N_{var} \times N_{var}$ correlation matrix $\boldsymbol{C}$. For each variable X_j ($j \leq N_{var}$), the Pearson correlation coefficient with variable X_i ($i \leq N_{var}$) is stored in $c_{i,j}$ (with special cases $c_{j,j} = 1$ and $c_{i,j} = c_{j,i}$). Subsequently, variable pairs with a correlation score exceeding the threshold value (τ_c) are identified and individually correlated with the response.

Here, the variable with the highest correlation with the response was maintained in the data set. In short, the procedure as shown by Algorithm 6.1 was applied. To assess the effect of correlation threshold selection, τ_c was set to range from 0.25 (high degree of removal) up to 0.95 (low degree of removal) with a step size equal to 0.05.

Algorithm 6.1: Correlation-based variable removal

```
Calculate correlation matrix C from dataframe D
FOR each element c in C
    IF element c is greater than or equal to correlation threshold τc
        Store unique variable-variable combination in an overall list L
    END if
END for
Sort list L according to decreasing correlation score
FOR each instance in list L
    Determine correlation of each variable with response
    Remove variable with lowest correlation from L and D
END for
```

Identification of irrelevant variables

The identification of irrelevant variables contrasts the straightforward correlation-based variable selection as it requires the development of a basic model to derive the importance scores of the incorporated variables. More specifically, variable importance was derived by developing CRFs and assessing the decrease in accuracy following permutation of the variable values, with higher scores being assigned to more important variables. As patterns and type of information differed among species, model-specific importance scores are divided by the highest obtained importance score and subsequently checked against a user-specified threshold (τ_i). All variables with a relative importance score below the threshold are consequently removed from the dataset (see Algorithm 6.2). To assess the effect of threshold selection, τ_i was set to range from 0 (low degree of removal) up to 0.5 (high degree of removal) with a step size equal to 0.05.

Algorithm 6.2: Importance-based variable removal

```
Develop basic model m
FOR each variable in D
    Derive variable importance scores from m
    Calculate relative variable importance
    IF relative importance is lower than threshold τi
        Remove variable from D
    END if
END for
```

6.2.4 Computation time and threshold selection

Improvement of data quality by eliminating instances and variables affects computation time in two ways: (1) it increases the time needed to prepare the data and (2) it potentially decreases the time needed to develop the individual model. To assess the consequences of abovementioned techniques, computation time was registered for the overall procedure including data preparation and model development as well as for the application of the model-developing algorithm. Hence, computation time for algorithm application reflects the average of all repetitions of 5-fold cross-validated models. In contrast, total time reflects the time needed to prepare the data and create repetitions of 5-fold cross-validated models, hence providing a single value per macrophyte.

For each technique, the effect of threshold selection on performance and time were visually assessed for the previously selected macrophytes (see Table 6.1), resulting in the suggestion of a single, technique-specific threshold value to be used. Subsequently, models were developed for all 58 macrophyte species, applying data preparation by combining the abovementioned techniques in the following order: (1) outlier removal, (2) false absence identification, (3) correlated variable removal and (4) irrelevant variable removal.

6.3 Results

6.3.1 Preliminary assessment

Range analysis for the hyperparameter *ntree* showed that model performance is only limitedly affected by the number of trees, with relatively stable performance along the studied range (Figure 6.2). Variability in performance increased with decreasing number of training instances (i.e. macrophyte prevalence), though hardly changed with increasing values of *ntree*. Therefore, a relatively low value (with respect to the default value *ntree* = 1000) can be selected to reduce the required calculation time. For instance, at *ntree* = 200 model performance is relatively stable (Figure 6.2), while reducing the model development time by 80 % due to the linear dependency between computation time and *ntree* (Stekhoven and Bühlmann, 2012). This value was considered for all further analyses and supported by the work of Oshiro *et al.* (2012), illustrating a limited increase of AUC above *ntree* = 200.

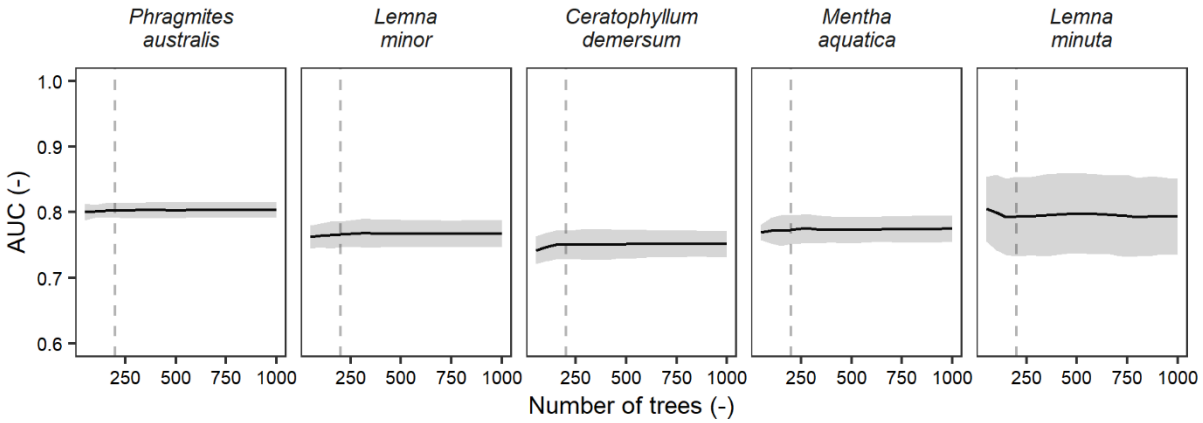

Figure 6.2: Model performance in function of the number of individual trees developed within the random forest. *Stability of performance (black line) can already be observed from 200 trees onward (dashed grey line), except for* L. minuta. *Variability in performance between folds (indicated as standard deviation in grey) is considered to be limited, though tends to increase as the number of training instances decreases, as illustrated by higher variability for* L. minuta *compared to* P. australis.

Similarly, including more repetitions to reduce overall variability in model performance already shows to be effective at low numbers of repetitions (Figure 6.3). For instance, after 7 repetitions the average performance related to *P. australis* and *L. minor* remains stable, while for *M. aquatica* and *L. minuta* some variability can still be observed. Variability in model performance among repetitions is higher for macrophytes with a lower number of training instances, and tends to remain relatively stable with increasing number of repetitions (Figure 6.3). Based on these observations, an overall guideline for number of repetitions within this study can be set on 10.

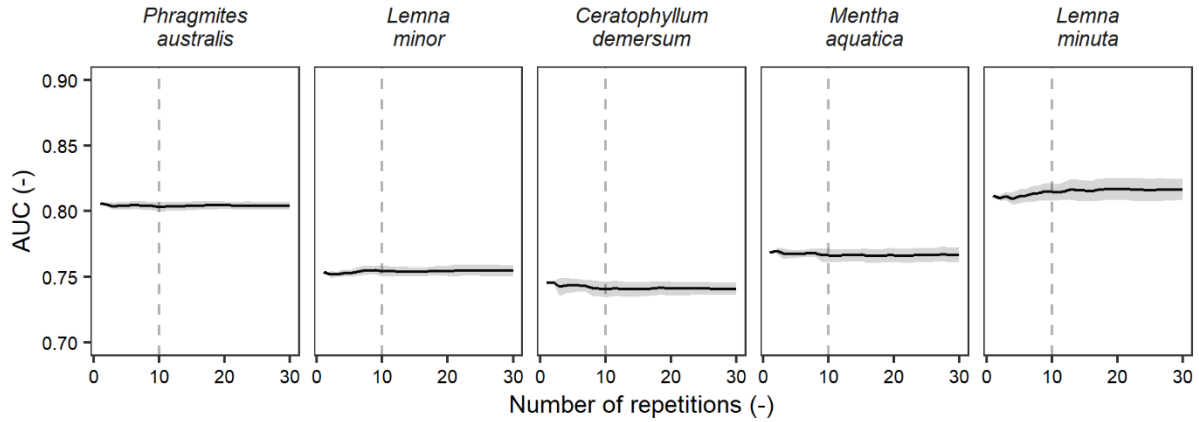

Figure 6.3: Model performance in function of the number of model repetitions. *Stability of performance (black line) can already be observed from 10 repetitions onward (dashed grey line). Overall, variability in average performance (indicated as standard deviation in grey) is limited, but tends to increase as the number of training instances decreases.*

6.3.2 Individual pre-processing

6.3.2.1 Instance-based removal

Excessively deviating instances were removed from the dataset for τ_o values ranging between 0 and 15. Generally, a decrease in model performance is obtained by outlier removal, yet shows to be relatively stable as soon as the most excessive outliers are removed (i.e. $15 < \tau_o < 10$). Further threshold reduction ($\tau_o \rightarrow 5$) considers more instances to be outliers, though causes only limited reduction of model performance for *P. australis*, *L. minor* and *C. demersum*, while models for *M. aquatica* and *L. minuta* already indicate a performance decrease when τ_o drops below 7. Overall, the effect of outlier removal on model performance is relatively limited, with a maximum decrease in AUC of 0.05 (*C. demersum*, see Figure 6.4).

In contrast, required computation time continuously decreases over the applied range for τ_o, showing a larger initial effect for *P. australis* compared to the other macrophytes. Moreover, time reduction shows a dependency on data availability with a gradual reduction for *P. australis* and a more abrupt reduction for *L. minuta* for τ_o-values smaller than 5. Similar patterns are observed for overall computation time, including the dependency on data availability (see Appendix, Figure C.6). For instance, an overall beneficial effect of outlier removal is observed for *P. australis*, while model development for *L. minuta* indicates to be negatively affected. In order to avoid an excessive performance decrease for low data-availability species, while already providing a 10-30 % reduction in computation time for high data-availability species, a threshold value of $\tau_o = 3$ can be derived, resulting in a removal of about 760 instances (see Appendix, Figure C.2).

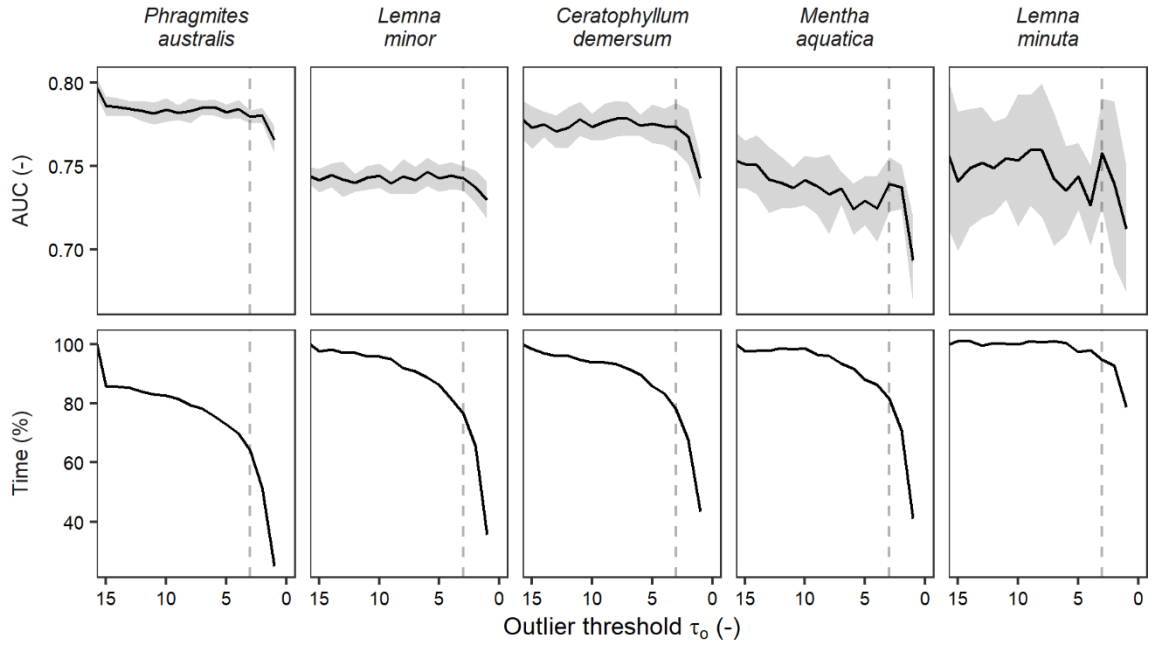

Figure 6.4: Effect of outlier-based instance removal on model performance and computation time. *Removal of outliers has, at first, a limited effect on performance and computation time (except for P. australis). A slight decrease in performance is observed when more deviating values are considered as outliers ($\tau_o \rightarrow 0$), while causing the required computation time to decrease. A visual trade-off between performance and computation time supports a threshold of $\tau_o = 3$ (dashed grey line). Performance analyses for all 58 species can be found in Appendix, Figure C.7 and Figure C.8.*

The removal of false absences provides a positive effect on model performance, with AUC values increasing as τ_a decreases, without reaching a plateau (Figure 6.5). As the threshold becomes more strict (i.e. $\tau_a \rightarrow 0$ %), performance keeps increasing up to net AUC improvements of 0.2 (*L. minor*). In general, patterns among macrophytes are relatively similar and show performance improvements for conservative threshold values (i.e. τ_a = 15 %), causing AUC scores to increase with about 0.05 (Figure 6.5). Similar analyses can be performed for the remaining macrophytes.

In contrast, computation time assessment indicates the existence of a species-specific tipping point for τ_a, below which computation time decreases drastically. These tipping points are related to overall data availability after false absence removal. For instance, data for *P. australis* originally represents about 1700 presences and around 2600 absences. As the threshold becomes stricter, more absences are removed, rising to 1000 at τ_a = 7 % and 2000 at τ_a = 0 % (see Appendix, Figure C.3), which results in only 1600 and 600 absences remaining, respectively. These absences are lower than the number of presences, which requires subsampling of the latter to create a balanced training set for model development. The resulting decrease in data size reduces the required computation time as less instances need to be classified.

A similar pattern is present in overall computation time, additionally showing an increase when data availability is too low to support faster model development (e.g. *L. minuta*) (see Appendix, Figure C.6). Consequently, any threshold value will affect model performance positively, yet selecting low values for τ_a (e.g. τ_a < 5 %) not only improves performance, but also causes high numbers of instances to be eliminated (up to 1200 instances, see Appendix, Figure C.3). A trade-off threshold value of τ_a = 5 % is suggested to avoid excessive removal.

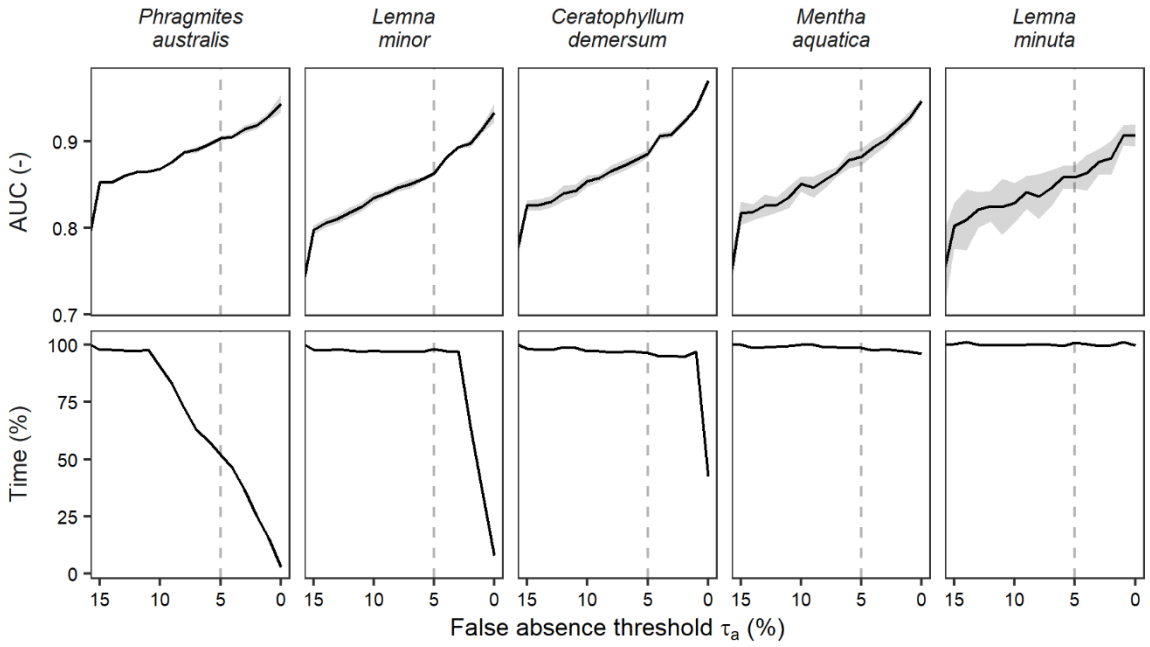

Figure 6.5: Effect of false-absence-based instance removal on model performance and computation time. *A continuous increase in performance is observed for each macrophyte as the threshold becomes more strict. In contrast, computation time remains relatively stable at first, while a sharp decrease is observed for some macrophytes as* τ_a *decreases. A trade-off between model performance, computation time and the consequences of removing too much ambiguous instances provides a compromise at* τ_a *= 5 % (dashed grey line). Performance analyses for all 58 species can be found in Appendix, Figure C.9 and Figure C.10.*

6.3.2.2 Variable-based removal

Correlated variables provide similar information, yet removal of these variables goes along with removal of information, illustrated by a decrease in model performance for decreasing correlation thresholds (Figure 6.6). At high threshold values (i.e. τ_c > 0.85), the reduction in performance remains relatively limited while at extremely low threshold values (i.e. τ_c < 0.40) a clear decrease in AUC values is observed due to the limited amount of shared information between limitedly-correlated variables.

Simultaneously, however, a gain in computation time is observed, following an overall dimensionality reduction within the search space caused by a decreased number of variables. The different plateaus observed within the time-specific graphs illustrate the inherent characteristics of the algorithm, selecting only a subset of all variables for each split within the tree. This number is based on the number of available variables and defined as $mtry = \sqrt{N_{var}}$, being rounded to the lower integer. Hence, as soon as N_{var} decreases sufficiently, *mtry* will drop with 1 unit, causing less variables to be selected and, consequently, less potential splitting points to be considered. Therefore, plateaus exist between each drop, as *mtry* does not change with every variable being removed.

Similar patterns are observed for overall computation time, showing generally faster data pre-processing and model development, though patterns become less clear as data availability decreases (see Appendix, Figure C.6). Threshold selection based on these results is not straightforward, yet was chosen at τ_c = 0.70 to avoid removal of more than 10 variables (see Appendix, Figure C.4).

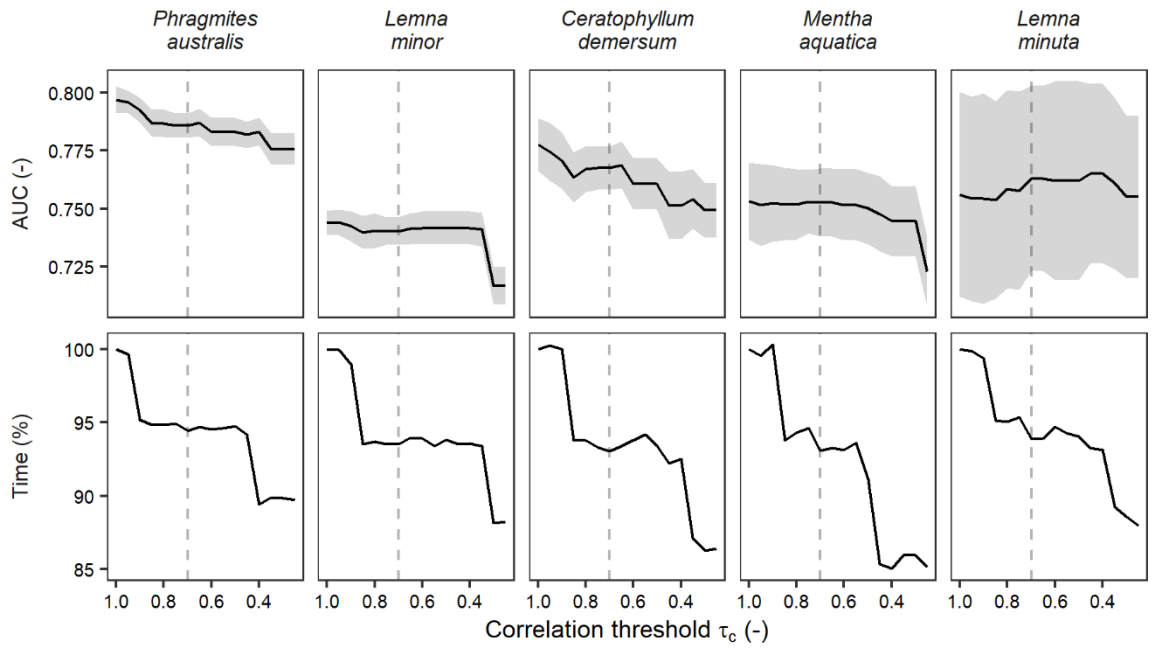

Figure 6.6: Effect of correlation-based variable removal on model performance and computation time. *Removal of correlated variables has no straightforward effect on performance, though a limited effect on performance and computation time is observed at first (τ_c > 0.9). Required computation time decreases with variable removal and illustrates the characteristic plateaus related with algorithm settings. Selection of an intermediate threshold value (i.e. τ_c = 0.7) considers variables with a relatively high correlation. Performance analyses for all 58 species can be found in Appendix, Figure C.11 and Figure C.12.*

A reduction in performance is also observed following the removal of irrelevant variables. As the required contribution of each variable increases (i.e. $\tau_i \rightarrow 50$ %) AUC values decrease to reach a macrophyte-specific plateau (Figure 6.7) caused by many variables being removed (see Appendix, Figure C.5). Nevertheless, at low threshold values (i.e. $\tau_i < 10$ %) model performance is hardly affected due to the removal of mostly irrelevant variables, while providing a minor decrease in computation time (1 to 6 %).

Similar to correlation-based variable removal, computation time decreases when more variables are eliminated, reaching a species-specific plateau. However, overall computation time tends to increase as the calculation of variable importance requires an additional model to be developed, being the main contributor to the overall required time (see Appendix, Figure C.6). Threshold setting at $\tau_i = 10$ % was supported by visual assessment of performance, computation time and number of variables being removed.

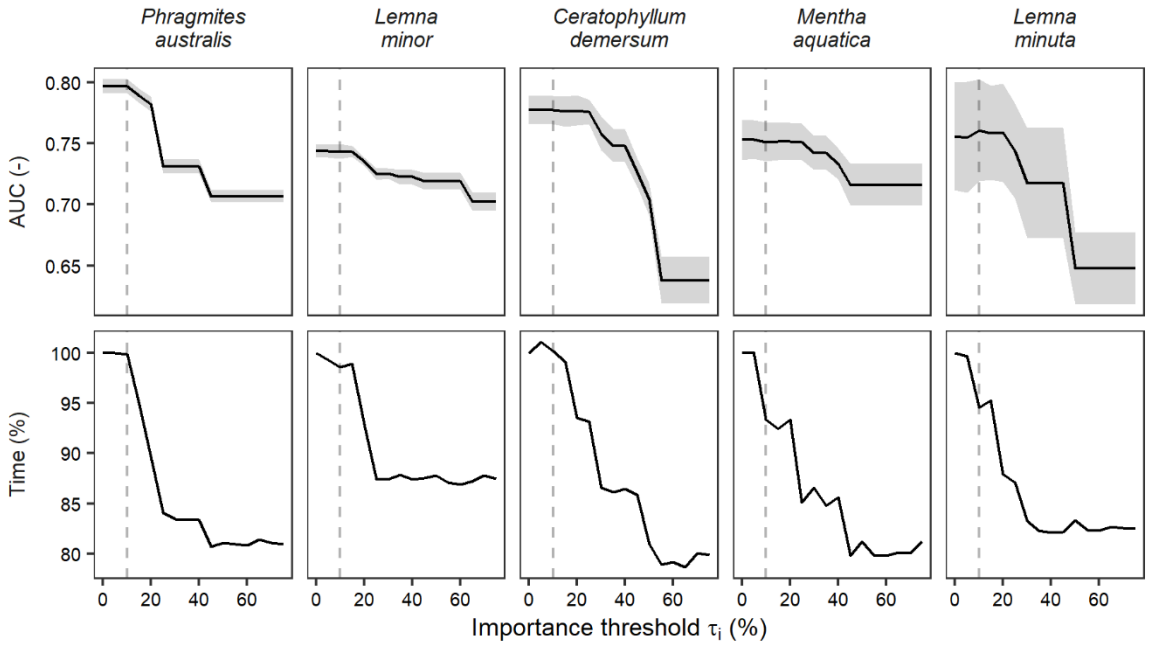

Figure 6.7: Effects of importance-based variable removal on model performance and computation time. *Removal of irrelevant variables has, at first, limited effect on performance and computation time. A clear decrease in performance can be observed as soon as relative importance scores exceed 20 %. In contrast, effects on computation time are already visible when removing the most irrelevant variables ($\tau_i < 15$ %). Threshold selection at τ_i (10 %, dashed grey line) illustrates the technique-specific trade-off between performance and speed. Performance analyses for all 58 species can be found in Appendix, Figure C.13 and Figure C.14.*

6.3.3 Overall pre-processing

Based on the results obtained for a selection of macrophytes, a set of thresholds was identified for overall data pre-processing regardless of the considered macrophytes (Table 6.2). Hence, these were used as general guidelines during further data pre-processing, while highlighting that future species-specific research can benefit from individual threshold analysis and setting. Here, however, the aim was to identify generally applicable threshold values rather than species-specific.

Table 6.2: Summary of technique-specific threshold values for data pre-processing. *Depicted threshold values were used during combinatory data pre-processing.*

Step	Threshold	Value
Outlier removal (-)	τ_o	3
False absence removal (%)	τ_a	5
Correlated variable removal (-)	τ_c	0.7
Irrelevant variable removal (%)	τ_i	10

Application of these thresholds supported a clear increase in model performance for the five selected macrophytes, showing an increase in AUC ranging between 0.799 ± 0.001 up to 0.848 ± 0.001 (*P. australis*) and 0.752 ± 0.003 up to 0.831 ± 0.004 (*M. aquatica*) (Figure 6.8). Similarly, data pre-processing showed to positively affect model performance for the majority of the 58 macrophytes, with AUC values after pre-processing being higher than the reference AUC values (Figure 6.9A). However, increased data pre-processing also affected the required computation time (Figure 6.9B), causing relative differences in computation time to be higher than the relative differences in AUC (Figure 6.9C).

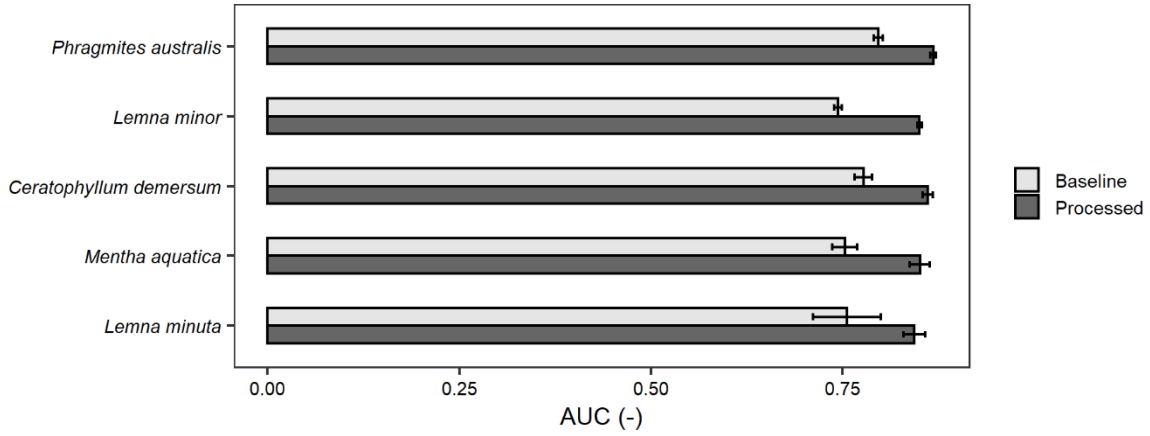

Figure 6.8: Effect of data pre-processing on model performance for a selection of five macrophytes, expressed as AUC. *An increase in performance is observed when data is pre-processed, contrasting baseline performance (light grey) versus performance following combinatory data pre-processing (dark grey). Error bars indicate the standard deviation over 10 repetitions of five-fold cross-validated models.*

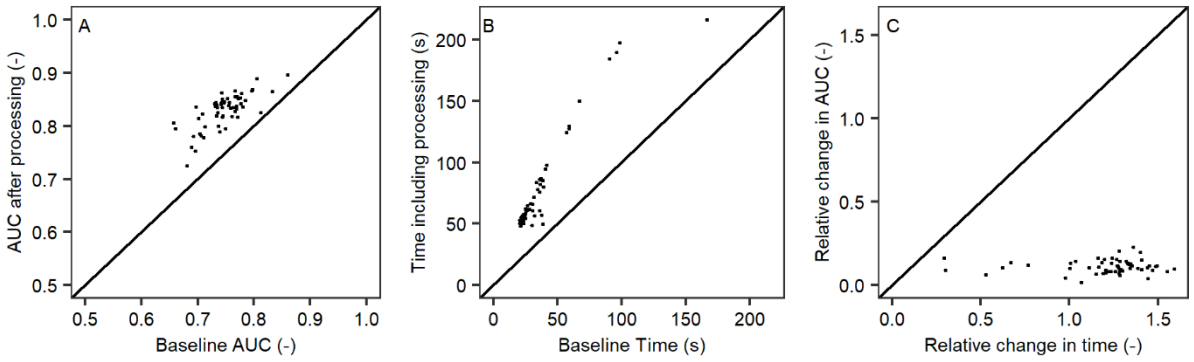

Figure 6.9: Effect of data pre-processing on performance and computation time for all considered macrophytes (N = 58). *Most models benefit from data pre-processing, yet require more computation time to improve data quality. A: Performance, expressed as AUC; B: Computation time, expressed in seconds; C: Relative change in performance versus relative change in computation time as part of a trade-off analysis. The diagonal black line indicates the agreement line with points above the line indicating an increase due to data cleaning (A, B) or higher relative change in performance compared to computation time (C).*

6.3.4 Final model evaluation

The resulting models were used to process a pseudo-independent dataset as a manner of testing the models' performance on external data. In general, external model performance was lower than internal model performance (Figure 6.10A), yet still provided acceptable models (AUC > 0.6). Additional processing of the test data (i.e. removal of potential false absences) increased external performance (Figure 6.10B) and showed to be slightly closer to internal model performance (Figure 6.10C).

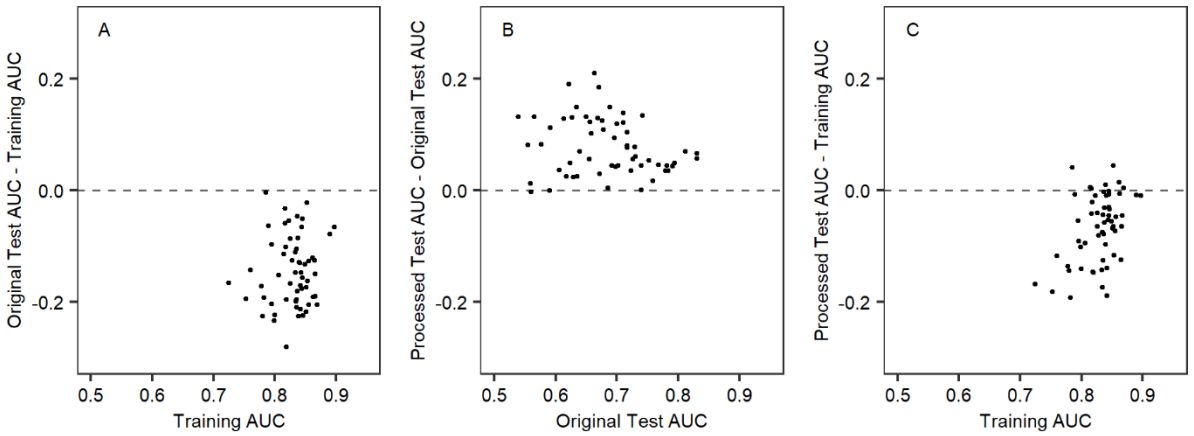

Figure 6.10: Model testing with external test set for all considered macrophytes (N = 58). *Testing was performed with two external datasets: (i) the original test set, (ii) the original test set devoid of false absences. Test performance was lower than internal performance, while processing the test set increased model performance. A: Difference between the original test performance and internal performance; B: Difference in performance between the processed and original test set and C: Difference in performance between the processed test set and the internal validation.*

6.4 Discussion

6.4.1 Data pre-processing affecting performance and speed

Generally, data cleaning clearly affected model performance, with AUC values declining along more stringent threshold values for three of the four pre-processing techniques, yet overall outperforming random classification (i.e. AUC > 0.5). Removal of outliers and variables (both correlated and irrelevant) showed to negatively affect model performance, depicting downward trends of AUC due to reduced data availability. Effects remained relatively limited, as illustrated by the removal of irrelevant variables causing the largest drop in AUC (i.e. from 0.78 to 0.64 for *C. demersum*), supporting the claim that random forests are relatively robust towards the inclusion of outliers and redundant variables (both correlated and irrelevant) (Breiman, 2001; Fox *et al.*, 2017; Vezza *et al.*, 2015). In contrast, improved model performance was observed following the identification and removal of potential false absences. More specifically, model performance showed a continuous increase in AUC along rising threshold levels (i.e. $\tau_a \rightarrow 0$), with highest performance scores being obtained when each instance within the assumed realised niche was removed from the background data.

The patterns obtained in this study comply with literature related to niche identification and predictor selection. For instance, Acevedo *et al.* (2012) showed that extending the environmental range made it easier to discriminate suitable from unsuitable habitats, thereby causing artificially increased AUC values. Hence, by decreasing the environmental range via outlier elimination, a drop in AUC scores is expected, which explains the obtained patterns in Figure 6.4. Similarly, Anderson and Raza (2010) applied a niche-corrected absence selection approach by excluding suitable conditions from the background data and observed an increase in model performance. By excluding these false absences, the distinction between suitable and unsuitable habitats was improved along with the support to obtain elevated AUC scores. Hence, by improving the discrimination within the observed environmental domain, a rise in AUC is expected, which supports the obtained performance increase in Figure 6.5.

In contrast, appropriate predictor selection supports an overall simplification of the observed environmental domain and, thus, model complexity. This niche simplification increases the model's transferability and application, as managers tend to request simple and understandable models (Bennetsen *et al.*, 2016). However, dimensionality reduction of the environmental domain rarely provides improved model performance, as predictors are either irrelevant or of limited importance within the observed domain. The exclusion of these predictors positively reduces model complexity, but negatively affects the combined explanatory power towards the observed variance in the response variable. Consequently, variable selection is expected to cause a decrease (or at least a stand-still) in performance, which clarifies the patterns in Figure 6.6 and Figure 6.7.

However, despite being applied and discussed in literature, it should remain clear that data pre-processing is not without consequences. Both instance and variable removal inherently affect data availability, species response curves and delineation of the occupied environmental domain. Preferably, only a fraction of the assessed environmental range is occupied by the species under consideration in order to distinguish between suitable and unsuitable habitats. However, as the extent of the considered biogeographical range is user-dependent and affects model performance conditions (Acevedo *et al.*, 2012; Anderson and Raza, 2010; Phillips *et al.*, 2009), care should be taken to delineate a reasonable domain. Moreover, the assumption underlying niche-based absence selection states that no unsuitable conditions exist within the observed realised niche, though extends to the idea that all relevant variables are observed and reported (Anderson and Raza, 2010). More specifically, it does not allow the presence of an unrecorded environmental variable or any biotic interaction to cause a species' absence, which supports model simplification and regularisation, but violates ecological theory.

Ultimately, model performance was improved through combinatory data pre-processing, following technique-specific threshold selection based on visual assessment of trends in performance, computation time and data characteristics. A general increase in model performance was observed, with net AUC improvements up to 0.2 and internal validation scores ranging between 0.7 and 0.9, supporting the claim that the improvement of data quality has potential beneficial effects on model performance. Slightly lower AUC scores were obtained when models were tested with an external data set (ranging between 0.54 and 0.83; average: 0.68 ± 0.07), due to the inclusion of false absences. Indeed, elimination of these absences significantly (Wilcoxon rank sum test; $W = 774.5$, $p < 0.001$) increased performance scores (ranging between 0.56 and 0.90; average: 0.76 ± 0.08) and suggested that remaining false absences might artificially deflate performance. This is especially the case when the external data is not a perfect subsample of the original distribution (e.g. rare species).

Lastly, data cleaning supported a decrease in the required computation time for model development for each pre-processing technique, while an overall increase in total computation time for combinatory pre-processing is obtained. Compared to the relative changes in performance, computation time changed drastically by implementing data cleaning, mostly showing an increase in pre-processing time and a decrease in model development time.

6.4.2 Implications for environmental research

Raw environmental data harbours an invaluable treasure of information, hidden in complex patterns and a significant amount of noise. Elimination of the latter simplifies pattern discovery and the development of species distribution hypotheses. The qualitative trade-off analyses performed here provided threshold values for the identification and elimination of outliers (τ_o = 3), false absences (τ_a = 5 %), correlated variables (τ_c = 0.7) and irrelevant variables (τ_i = 10 %). Despite frequent application within correlative ecological modelling, threshold values are only limitedly reported and often case-specific, underlining the need for a solid conceptual framework to govern sound and comparable results and conclusions to support decision-making (Kotsiantis *et al.*, 2006; Zhang *et al.*, 2003).

Unfortunately, data collection and cleaning remain expensive steps within species distribution studies (Zhang *et al.*, 2003). To start, data collection by means of field campaigns is time-, energy- and budget-intensive, causing researchers to refrain from data removal and data sharing, which increases the need for thorough data cleaning (Catalano *et al.*, 2019). Recent movements towards open data and uniform data bases (e.g. Global Biodiversity Information Facility, GBIF) have eased the process of gathering occurrence information, thereby causing an exponential growth in occurrence-based modelling of habitat suitability and species distributions (Peterson *et al.*, 2015). Yet, the available data is to be used with care as the provided quality is subject to the preferences of the original owner of the data (Maldonado *et al.*, 2015), causing data reliability to become an additional aspect to be considered within correlative habitat suitability and species distribution modelling. For instance, herbaria and museums are increasingly improving data availability by digitising their collections, though these observations often bias results as they lack detailed georeferencing (Maldonado *et al.*, 2015; Peterson *et al.*, 2015). In addition, due to the high variety in data quality, data cleaning can take up to 80 % of all time spent on a research project (Zhang *et al.*, 2003). Even when automated, further tuning remains necessary to find the appropriate threshold values.

Here, the selected techniques have been tuned manually to act as a filter for the data to be used, while they provide the opportunity to be included in the model development algorithm and act as wrapper functions with tuneable hyperparameters (e.g. Boets *et al.* (2013a), Gobeyn *et al.* (2017)). Moreover, alternative approaches do exist, including visual outlier identification (Gobeyn *et al.*, 2017), distance-based pseudo-absence selection, input variable selection by means of Genetic Algorithms (D'Heygere *et al.*, 2003; Gobeyn *et al.*, 2017), variable transformation (Kotsiantis *et al.*, 2006) and variable construction (Kotsiantis *et al.*, 2006). Each of these techniques includes some kind of user-dependent threshold selection and influences model performance and output (including decision-making) differently. This underlines the need for a well-developed framework to support sound model development.

6.4.3 Contribution to the study objective

The aim of this chapter was to assess the effects of technique-specific threshold selection on model performance and the required computation time in order to provide guidelines for further pre-processing of the adopted Limnodata Neerlandica. Throughout the chapter, threshold values were altered to infer their effect on model performance and to allow a trade-off between model performance, computation time and data loss. By considering these ranges, a more pronounced basis was created to bring forward a set of threshold values for supporting after-imputation data cleaning within the overall study objective (see Section 1.2.1). Similar to Chapter 5, it should remain clear that this chapter contributes mostly to the overall study objective, while providing suggestions for application outside the considered framework. More specifically, it is recommended to perform similar analyses with different combinations of environmental variables and species occurrences to support empirical threshold selection.

The chapter complies to the recommendation of performing data pre-processing prior to data-driven model development in order to eliminate noise within publicly available data (Maldonado *et al.*, 2015). It was expected that noise was present in the Limnodata Neerlandica, as data was collected by various companies and institutions over a period of thirty years (see Section 4.2.1). More specifically, this noise was expected to be present in the instances (i.e. extremely deviation values, recording of false absences) and among the variables (i.e. correlations and non-influential variables), with a potential to negatively affect model performance (Murphy *et al.*, 2010). In literature, noise elimination through data pre-processing is often done in a partial and subjective manner (e.g. Forio *et al.* (2018), Fox *et al.* (2017), Gobeyn *et al.* (2017)), though deserves more scrutiny due to its negative effect on data availability.

In general, the removal of noise (outliers, false absences, correlated and irrelevant variables) supported the expected changes in model performance, although three out of four methods caused a decrease in the performance metric score (see Section 6.3.2). Only the removal of false absences affected model performance positively, mainly due to a clearer delineation of the realised niche. Due to the performed range assessment, threshold values for the pre-processing of the imputed Limnodata Neerlandica could be defined via a visual trade-off between model performance, computation time and data availability, resulting in thresholds for the elimination of outliers ($\tau_o = 3$), false absences ($\tau_a = 5$ %), correlated variables ($\tau_c = 0.7$) and irrelevant variables ($\tau_i = 10$ %). By performing such a visual trade-off, a certain degree of subjectivity is introduced, yet this is considered to be lower than simply adopting thresholds from similar studies. More importantly, the implementation of these pre-processing thresholds creates species-specific data sets, which support the construction of qualitative models to describe the abiotic suitability of wetland habitats for specific aquatic macrophytes.

6.5 Conclusion

Occurrence data contain valuable information on species distribution patterns and dynamics, but require data cleaning prior to pattern inference. During cleaning, data is unavoidably lost as environmental domains become more strictly delineated. Identification and elimination of outliers and variables that are correlated or irrelevant inherently increase potential overlap of presence and background domains, while discarding potential false absences supports the identification of more distinct (yet less detailed) environmental niches. Accordingly, a decrease or increase in model performance is observed whenever the environmental domains of presences and absences are characterised by respectively more or less relative overlap due to data quality improvement. In contrast, a decrease in computation time required for model development is observed for each type of data cleaning, with inclusion of the data pre-processing step causing overall computation time to be both lower and higher than without data pre-processing, depending on the applied technique. A visual trade-off analysis of performance and computation time, supplemented with the effects of threshold selection on the sample size or dimensionality of the data, identifies thresholds for the elimination of outliers (τ_o = 3), false absences (τ_a = 5 %), correlated variables (τ_c = 0.7) and irrelevant variables (τ_i = 10 %), while supporting improved model performance following combinatory data pre-processing. The increased data quality and resulting decreased model complexity underline the added value of data pre-processing within the framework of species distribution modelling and model transferability.

7

Abiotic habitat suitability models to assess restoration potential and invasion vulnerability[5]

Highlights

- Only a fraction of the suitable abiotic habitats is occupied by macrophytes
- Key variables are temperature, pH, nitrate, ammonium and oxygen
- Managing key variables impacts habitat suitability more than business-as-usual
- Models are able to identify locations with high invasion potential

[5] This chapter is based on Van Echelpoel, W.; Forio, M. A. E. and Goethals, P. L. M. (in preparation) Abiotic habitat suitability models as first-level assessment for restoration potential and invasion vulnerability

Abstract

Macrophytes have a steering role in ecosystem functioning, yet their presence is affected by a myriad of physical, chemical and biological variables. Improving and safeguarding macrophyte-influenced ecosystem services requires identification and management of suitable habitats. First-level habitat suitability scores were defined by linking abiotic conditions with presence/absence data for 58 macrophyte species by means of conditional random forests. Developed models showed good discriminative and classification power, with final AUC (Area Under the receiver operating characteristic Curve) values between 0.846 ± 0.008 and 0.888 ± 0.002, while sensitivity and specificity ranged between 0.736 ± 0.008 and 0.796 ± 0.003 and between 0.738 ± 0.007 and 0.791 ± 0.002, respectively. Temperature, nitrate, oxygen, ammonium and pH were major abiotic habitat descriptors and affected habitat suitability in a similar, yet species-specific way. In general, suitability scores increased along rising temperature and pH values, followed by a drop at high pH levels (> 8.5). In contrast, a negative effect of rising nitrate and ammonium levels on habitat suitability occurred, confirming the anticipated positive impact of pollution reduction on macrophyte presence. Management aiming at optimising nitrate-nitrogen (0.5 $mg{\cdot}L^{-1}$ up to 1.5 $mg{\cdot}L^{-1}$), oxygen (4 $mg{\cdot}L^{-1}$ up to 7 $mg{\cdot}L^{-1}$), ammonium-nitrogen (0.3 $mg{\cdot}L^{-1}$ up to 0.5 $mg{\cdot}L^{-1}$) and pH (7 up to 8.5) will positively impact the chances for macrophyte survival. Historically, species prevalence has been increasing and is generally characterised by a lag between predicted and observed presence, though this trend is expected to continue. Yet, improved abiotic conditions can indirectly threaten native macrophyte species when also habitat suitability for invasive alien species increases. Similar patterns were observed for the native *Lemna minor* and alien *Lemna minuta*, requiring further quantification of physiological processes via laboratory experiments to elucidate actual field effects.

7.1 Setting the scene

In Chapter 2 it became clear that the conservation of ecosystem structure and functioning within wetlands should focus on macrophytes to benefit from their capacity to compartmentalise the prevailing habitat. Identifying optimal conditions and strategies underlies management success and is highly supported by the development of habitat suitability models (HSMs), which often rely on publicly available data. Chapter 5 and Chapter 6 highlighted some opportunities to improve data quality and thereby provided the data-related foundation of this chapter. Here, the application of HSMs for inferring optimal habitats for macrophyte presence is introduced and discussed within a conservation framework.

Macrophyte management represents a challenging endeavour as their presence is affected by a combination of geomorphological, hydrological, chemical and biological conditions (Bakker *et al.*, 2013; Bornette and Puijalon, 2011). For instance, historic eutrophication caused drastic decreases in macrophyte stocks due to the proliferation of phytoplankton, thereby increasing turbidity, toxic compounds and oxygen fluctuations (Scheffer *et al.*, 2001; Scheffer *et al.*, 1993b). Even with improved abiotic conditions and reduced phytoplankton competition, no straightforward restoration path to the initial biotic conditions exists. This multitude of potential pathways is caused by a myriad of biotic processes, including (propagule) dispersal, seed bank composition and presence of opportunistic species (Bakker *et al.*, 2013; Scheffer *et al.*, 1993b).

In addition, increasing globalisation amplifies the pressure of invasive alien species towards aquatic systems, leading to physical, chemical and biological habitat changes caused by intentional and unintentional introductions (Richter *et al.*, 2003; Sala *et al.*, 2000). Hence, conservation and improvement of native macrophyte habitats require the identification of (i) habitats suitable for supporting macrophyte presence, (ii) habitats vulnerable to invasion, distinguishing between sites with and without native species being present and (iii) habitats that require optimisation of their abiotic conditions and, if possible, which variable(s) to focus on. HSMs can provide such information, but with the important side note that due to their correlative nature, no undisputable conclusions on causality can be inferred.

Within this chapter, conditional random forests (CRFs) are developed and optimised to derive habitat suitability for a selection of macrophyte species. The aim is to combine ecological restoration and invasive alien species management by defining the effect of species-specific key variables on habitat suitability and elaborating on management options to optimise abiotic conditions. By tackling these issues, an answer is provided to RQ2.2, as defined in Chapter 1. Hence, this chapter concludes with a statement on which variables generally affect habitat suitability and how management can help with reaching optimal conditions.

7.2 Materials and methods

7.2.1 Characterisation of the data and modelling technique

Data within the Limnodata Neerlandica was subsampled to contain spatiotemporally referenced observations of macrophytes and the prevailing physicochemical conditions (see Chapter 4), providing information on 4344 instances, 176 variables and 576 macrophytes. Data pre-processing was performed as outlined in Chapter 6, following (i) missing data imputation, (ii) macrophyte selection, (iii) outlier removal, (iv) false absence removal, (v) correlated variable removal and (vi) irrelevant variable removal. Consequently, for each macrophyte, a specific data set was created due to the pre-processing being partially macrophyte-specific.

Ultimately, data for 58 macrophytes were available (see Appendix, Table D.1 and Figure D.1), yet only a subset of five macrophytes with varying prevalence level, growth form and origin will be highlighted in more detail (see also Chapter 6, Table 6.1): *Phragmites australis* (55 %), *Lemna minor* (44 %), *Ceratophyllum demersum* (29 %), *Mentha aquatica* (18 %) and *Lemna minuta* (5 %). Species prevalence within these data sets is higher than reported in Table 6.1 and intrinsically linked to the removal of false absences during data pre-processing. Additional R-packages for this chapter were *party* and *PresenceAbsence* (Freeman and Moisen, 2008a; Stekhoven, 2013).

Conditional random forests were developed to link macrophyte occurrence with the prevailing abiotic conditions, starting at default hyperparameter values, except for *ntree*, which was set at 200 (see Section 6.3.1). Subsequently, hyperparameter settings were optimised by means of randomly sampling the initial global search space, followed by an iterative optimisation within a local search space. Evaluation of model performance was done with AUC, Sn and Sp (see Section 3.4.2.1) and contrasted with species-specific null models. Finally, species-specific variable importance scores were determined via the developed models (Model Improvement Ratios; MIRs) and used for partial dependence analysis. A detailed description of the methodology can be found in Chapter 4.

7.2.2 Model application

A positive temporal trend in both habitat suitability and macrophyte occurrence was expected due to improved management and dispersal. Optimised models were applied to the original (imputed) data set to infer macrophyte-specific habitat suitability scores for all sampled sites. Discretisation of the Habitat Suitability Index (HSI) scores followed threshold identification via minimising the absolute sensitivity-specificity difference and subsequent temporal grouping to derive annual prevalence (predicted number of suitable sites divided by the total number of sites). Observed and predicted annual prevalence were compared to infer (i) the temporal trend of macrophyte prevalence and (ii) the potential macrophyte presence.

To mimic the potential effects of changing abiotic conditions on habitat suitability and illustrate the value of the constructed species-specific models towards management, six scenarios were developed. These scenarios represent three starting conditions (average, extreme and nutrient enrichment) and two management options (business-as-usual and focus on key variables), as mentioned in Table 7.1 and summarised in Table 7.2. The starting conditions were based on the observed environmental conditions in 2010 due to a lack of sufficient data from subsequent years. Moreover, observations were limited to the months April until September to limit seasonal bias within the temporal trends.

For each variable, the mean ($\bar{\boldsymbol{x}}$) and standard deviation ($\boldsymbol{s}$) were estimated (see Appendix, Table D.2) and used as a statistical basis for determining the three different starting conditions. First, the variable means were adopted when the starting conditions were defined to represent the average situation ($\bar{\boldsymbol{x}}$; 'AVG' scenarios). Secondly, nutrient-related variable means were increased with two times the standard deviation to reflect eutrophic sites, representing the nutrient-enriched situation ($\bar{\boldsymbol{x}}$ for non-nutrient variables and $\bar{\boldsymbol{x}} + 2 \cdot \boldsymbol{s}$ for nutrient variables; 'NUT' scenarios). Thirdly, variable means were increased with two times the standard deviation to reflect highly polluted sites, representing the extreme situation ($\bar{\boldsymbol{x}} + 2 \cdot \boldsymbol{s}$; 'EXT' scenarios). Several exceptions were considered in the latter, as pollution is reflected differently within the included environmental variables. More specifically, temperature and pH were not changed (i.e. $\bar{\boldsymbol{x}}$) and oxygen (saturation) was decreased instead of increased (i.e. $\bar{\boldsymbol{x}} - 2 \cdot \boldsymbol{s}$). Actual values can be found in Appendix, Table D.3.

For each variable, specific end points were defined depending on the performed management activities. First, variable-specific temporal trends were used for deriving the average change rates for each individual variable, reflecting the business-as-usual situation ('BAU' scenarios). Secondly, partial dependence plots were used for identifying the key habitat descriptors and their associated optimal conditions, reflecting management with a focus on the main habitat descriptors ('KEY' scenarios). For these key variables, an exponential temporal pattern was assumed, while all remaining variables were assumed to follow the temporal pattern as defined in the BAU scenario. The actual values can be found in Appendix, Table D.3.

Table 7.1: Assignment of scenario-specific codes. *Business-as-usual management relies on the continuation of variable-specific historical trends, while management focusing on key variables considers the optimal values of partial dependence plot as management endpoints. Starting point conditions are derived from observation data gathered in 2010. A more detailed description of each scenario can be found in Table 7.2.*

	Average conditions	**Extreme conditions**	**Nutrient enrichment**
Business-as-usual	AVG-BAU	EXT-BAU	NUT-BAU
Key variables	AVG-KEY	EXT-KEY	NUT-KEY

Table 7.2: Characterisation of management scenarios under different starting conditions. *Information extends the codes mentioned in Table 7.1.*

Code	Description
AVG-BAU	Baseline starting point with business-as-usual management. Starting point of each variable represents the average value observed in 2010. Management entails no alterations towards the previous period, hence the same temporal trend is assumed. Trends were derived by fitting variable-specific linear models to the temporal data (see Figure D.3).
AVG-KEY	Baseline starting point with management focusing on key variables. Starting point of each variable represents the average value observed in 2010. Management entails variable-specific procedures being solely applied to the five key variables, with endpoints derived from the partial dependence plots (see further).
EXT-BAU	Extreme starting point with business-as-usual management. Starting point of each variable represents the mean observed in 2010, supplemented with two times the standard deviation ($\bar{x} + 2 \cdot s$). Variable-specific exceptions were considered, depending on the included variables. Management entails no alterations towards the previous period, hence the same temporal trend is assumed. Trends were derived by fitting variable-specific linear models to the temporal data (see Figure D.3).
EXT-KEY	Extreme starting point with management focusing on key variables. Starting point of each variable represents the mean observed in 2010, supplemented with two times the standard deviation ($\bar{x} + 2 \cdot s$). Management entails variable-specific procedures being applied to the five key variables, with endpoints derived from the partial dependence plots (see further).
NUT-BAU	Nutrient enrichment with business-as-usual management Starting point of each nutrient variable represents the mean observed in 2010, supplemented with two times the standard deviation ($\bar{x} + 2 \cdot s$). Variable-specific exceptions were considered, depending on the included variables. Management entails no alterations towards the previous period, hence the same temporal trend is assumed. Trends were derived by fitting variable-specific linear models to the temporal data (see Figure D.3).
NUT-KEY	Nutrient enrichment with management focusing on key variables. Starting point of each nutrient variable represents the mean observed in 2010, supplemented with two times the standard deviation ($\bar{x} + 2 \cdot s$). For all other variables, the starting point was represented by the average value observed in 2010. Management entails variable-specific procedures being solely applied to the five key variables, with endpoints being defined by the partial dependence plots (see further).

The effects of the different scenarios on habitat suitability were subsequently assessed by applying the optimised macrophyte-specific models and deriving the suitability index. It should be noted that these scenarios were not developed to closely represent actual natural conditions and trends, but rather to illustrate the potential usage of the constructed models to assess scenario outcomes in function of the considered starting conditions. The obtained outcomes are meant to illustrate how management decisions can be steered by prevailing abiotic conditions.

Finally, the developed models were considered to contrast habitat preferences between two congeneric species. More specifically, occurrence observations of the native *Lemna minor* and the alien *L. minuta* (see Box 7.1) were confronted with predictions to determine (i) the ability of conditional random forest to identify suitable habitats for both *Lemna* spp. and (ii) whether the majority of the sites were more likely to support *L. minor* than *L. minuta*. It should be noted that the results have to be interpreted with care, as (i) data covered almost 30 years of sampling, (ii) pseudo-absences were used and (iii) *L. minuta* was relatively recently introduced (thus expected to violate the equilibrium assumption (Gallien *et al.*, 2012)).

Box 7.1: Selection of *Lemna minor* and *Lemna minuta*

The freshwater system that is considered as baseline throughout this work is characterised by slow-flowing water and elevated nutrient conditions (see Section 1.2.1). These conditions strongly support the presence of floating macrophytes, including the free-floating duckweed species (Bakker *et al.*, 2013; Zhang *et al.*, 2017). Among these duckweeds, *Lemna minor* frequently occurs in European surface waters, while *Lemna minuta* originates from North and South America and has reached a widespread status throughout Europe (Hussner, 2012). *L. minor* and *L. minuta* are morphologically similar and are often reported in the same locations, though their habitat preferences are not necessarily identical.

The development of species-specific models allows for distinguishing habitat preferences between these congeneric species and identifying the consequences of management on species-specific habitat suitability. Moreover, it can be used as an early-warning tool to locate sites with significantly higher HSI scores for the alien species compared to the native species. However, such applications merely illustrate preferences and suitability scores, while actual management decisions on avoiding species presence are to be made by the user.

7.3 Results

7.3.1 Model performance and optimisation

Hyperparameter optimisation provided a selection of species-specific settings, depicting an overall increase in *ntree* and decrease in *mtry*, when compared to the baseline settings (i.e. 200 and $\sqrt{N_{Var}}$, respectively), see Table 7.3. These settings were used to perform all subsequent analyses. Differences between internal and external validation were observed to be minimal (see Table 7.4), indicating that overfitting within the developed models hardly occurred. Surprisingly, differences in performance between the baseline and optimised models were often small (Table 7.4), suggesting a limited influence of hyperparameter tuning within this framework. Moreover, due to specifying *nsplit* and *nleaf* relative to the number of training instances (instead of absolute, see Section 4.2.3.3), model performance tended to be slightly lower when applying the optimal hyperparameter set. More specifically, it restricted the size of each individual tree within the random forest, thereby reducing complexity at the expense of performance.

External validation of species-specific models with pseudo-independent data indicated good model performance, with AUC values ranging between 0.85 ± 0.02 (*Lemna minuta*) and 0.888 ± 0.005 (*Ceratophyllum demersum*). Sensitivity and specificity were generally lower than AUC scores, but followed a similar pattern by ranging between 0.74 ± 0.03 *(L. minuta*) and 0.796 ± 0.008 *(C. demersum*) and between 0.74 ± 0.02 (*L. minuta*) and 0.791 ± 0.007 (*C. demersum*), respectively (Table 7.4). All models greatly outperformed null models, with 95-percentile scores between 0.596 (*Phragmites australis*) and 0.653 (*L. minuta*) for AUC, between 0.561 (*P. australis*) and 0.604 (*L. minuta*) for sensitivity and between 0.560 (*L. minor*) and 0.605 (*L. minuta*) for specificity (Table 7.4).

Table 7.3: Selected hyperparameter settings for conditional random forest development linking species occurrence to abiotic conditions. *Four hyperparameters were varied during the optimisation process, being* ntree *(number of individual models to be developed in the ensemble),* mtry *(number of variables to be considered for each split within the tree),* nsplit *(minimum fraction of instances in a node in order to be considered for splitting) and* nleaf *(minimum fraction of instances in a terminal node in order to be kept).*

Macrophyte	*ntree*	*mtry*	*nsplit*	*nleaf*
Phragmites australis	1540	2	0.04	0.01
Lemna minor	1890	2	0.09	0.01
Ceratophyllum demersum	1690	2	0.09	0.01
Mentha aquatica	1290	2	0.04	0.01
Lemna minuta	1040	2	0.09	0.01

Table 7.4: Overview of performance scores for a selection of macrophytes. *Null models were developed with permuted data and 95-percentiles were derived from 1000 models. The baseline model applies default hyperparameter values, while the optimised model makes use of adapted hyperparameter settings (see Table 7.3). Both model types were evaluated internally (cross-validation) and reported as* Baseline *and* Optimised. *The optimised model was also evaluated externally with a pseudo-independent test set (10 % of original data), being reported as* Evaluation. *Performance is described by Area under the Receiver Operating Characteristic Curve (AUC), sensitivity (Sn) and specificity (Sp), and rounded to three digits.*

Macrophyte	AUC	Sn	Sp
Phragmites australis			
Null model (P_{95})	0.596	0.561	0.562
Baseline	0.874 ± 0.003	0.783 ± 0.007	0.782 ± 0.007
Optimised	0.863 ± 0.003	0.772 ± 0.007	0.772 ± 0.006
Evaluation	0.850 ± 0.002	0.756 ± 0.003	0.754 ± 0.004
Lemna minor			
Null model (P_{95})	0.596	0.561	0.560
Baseline	0.839 ± 0.005	0.751 ± 0.007	0.748 ± 0.008
Optimised	0.823 ± 0.004	0.743 ± 0.006	0.744 ± 0.006
Evaluation	0.851 ± 0.003	0.753 ± 0.005	0.755 ± 0.005
Ceratophyllum demersum			
Null model (P_{95})	0.621	0.577	0.577
Baseline	0.861 ± 0.006	0.770 ± 0.009	0.770 ± 0.010
Optimised	0.854 ± 0.006	0.768 ± 0.009	0.765 ± 0.008
Evaluation	0.888 ± 0.005	0.796 ± 0.008	0.791 ± 0.007
Mentha aquatica			
Null model (P_{95})	0.609	0.569	0.568
Baseline	0.857 ± 0.008	0.769 ± 0.007	0.768 ± 0.009
Optimised	0.862 ± 0.008	0.778 ± 0.009	0.776 ± 0.010
Evaluation	0.856 ± 0.007	0.757 ± 0.011	0.756 ± 0.010
Lemna minuta			
Null model (P_{95})	0.653	0.604	0.605
Baseline	0.842 ± 0.026	0.764 ± 0.027	0.753 ± 0.019
Optimised	0.854 ± 0.025	0.774 ± 0.020	0.766 ± 0.018
Evaluation	0.846 ± 0.024	0.736 ± 0.025	0.738 ± 0.023

7.3.2 Variable importance

The importance of environmental variables to describe the occupied habitats varied among species and showed to be relatively high for temperature and nitrate (see Appendix, Figure D.2). Within the selected subset of macrophyte species, both variables were among the five most informative variables, with MIRs ranging between 1.00 (s < 0.01) (*P. australis*) and 0.9 ± 0.2 (*L. minuta*) for temperature and between 1.00 (s < 0.01) (*L. minor*) and 0.4 ± 0.1 (*L. minuta*) for nitrate (Figure 7.1). Inclusion of chlorophyll *a* during model development tended to be beneficial for *L. minor*, *C. demersum* and *L. minuta*, while models for *P. australis* and *M. aquatica* were more affected by ammonium and pH. Oxygen supported habitat description for both *Lemna* spp., while sulphate provided additional explanation for *L. minuta* and *C. demersum* (Figure 7.1).

Additional informative variables for these macrophytes included chloride (*P. australis*), potassium (*L. minor*), Kjeldahl-nitrogen (*C. demersum*) and total phosphorus (*M. aquatica*) as depicted in Figure 7.1. An overview of variable importance for all considered macrophytes (58 species) is provided in Appendix (Figure D.2), illustrating the dominance of both temperature and nitrate over other variables. On average (i.e. over all 58 species), temperature was characterised by the highest MIR (0.7 ± 0.3), followed by nitrate (0.5 ± 0.3), oxygen (0.3 ± 0.3), ammonium (0.3 ± 0.2) and pH (0.3 ± 0.2).

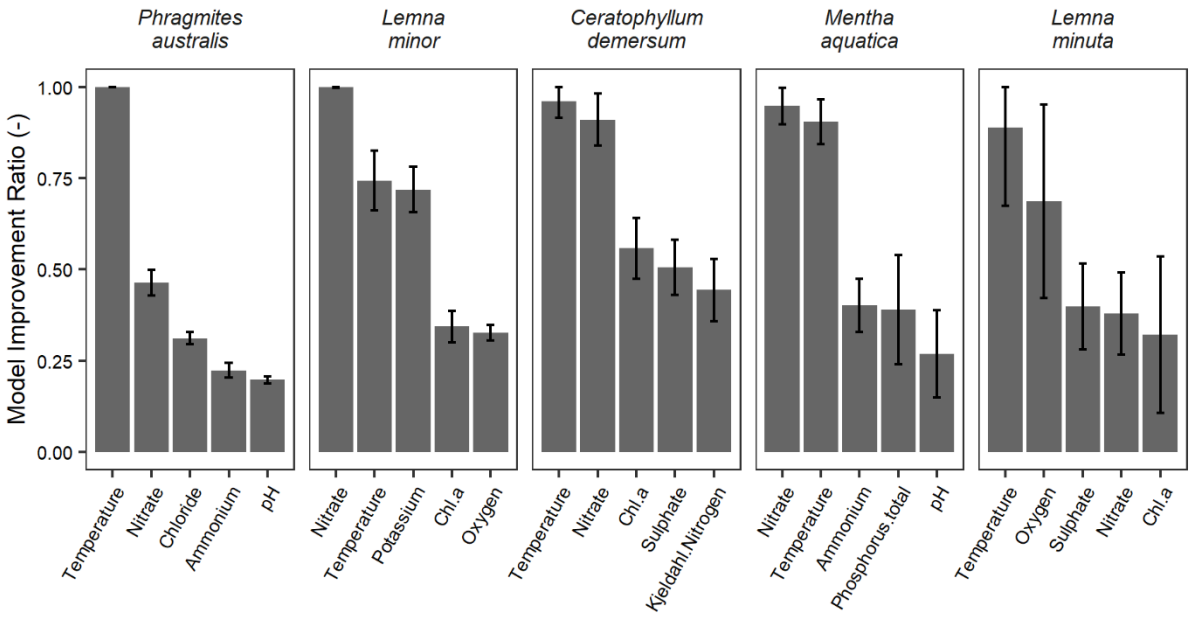

Figure 7.1: Variable importance of the five most informative variables for a selection of macrophytes. *Variable importance is expressed as Model Improvement Ratio (MIR), describing the relative importance of a variable with respect to the most informative variable. Temperature and nitrate recur for each macrophyte with either one as the most influential variable, while highly equal scores between both variables are obtained for* C. demersum *and* M. aquatica. *Vertical black lines indicate the standard deviation on the calculated MIRs.*

Changes in temperature, nitrate, oxygen, ammonium and pH showed a clear impact on the habitat suitability index (HSI) of the selected macrophyte species, although the magnitude of the effect declined along decreasing average variable importance (Figure 7.2). Higher temperatures tended to have a positive effect on habitat suitability for each macrophyte, with the highest increase in average HSI for *P. australis* (from 0.240 ± 0.008 up to 0.593 ± 0.004). Steep improvements in habitat suitability mainly occurred between 12 and 17 °C, while reaching an optimum around 20 °C (Figure 7.2).

Analogous patterns were observed for the remaining four variables, showing an overall negative effect on HSI when aquatic conditions were becoming too extreme. For instance, an optimal pH range was observed between 7 and 8.5 with lower HSI scores towards both extremes, while also oxygen indicated higher habitat suitability when concentrations ranged between 2 mg·L^{-1} and 7 mg·L^{-1} (Figure 7.2). Similarly, nitrate and ammonium showed a clear hormesis effect on habitat suitability as HSI scores were highest at concentrations above complete absence (i.e. 0 mg·L^{-1}) and below the observed extremes. More specifically, optimal conditions were slightly above zero (around 0.5 mg·L^{-1} for nitrate-N and 0.2 mg·L^{-1} for ammonium-N) and indicated generally suboptimal conditions at higher levels, which illustrates the potential negative effects of fertiliser run-off and wastewater discharge on macrophyte presence.

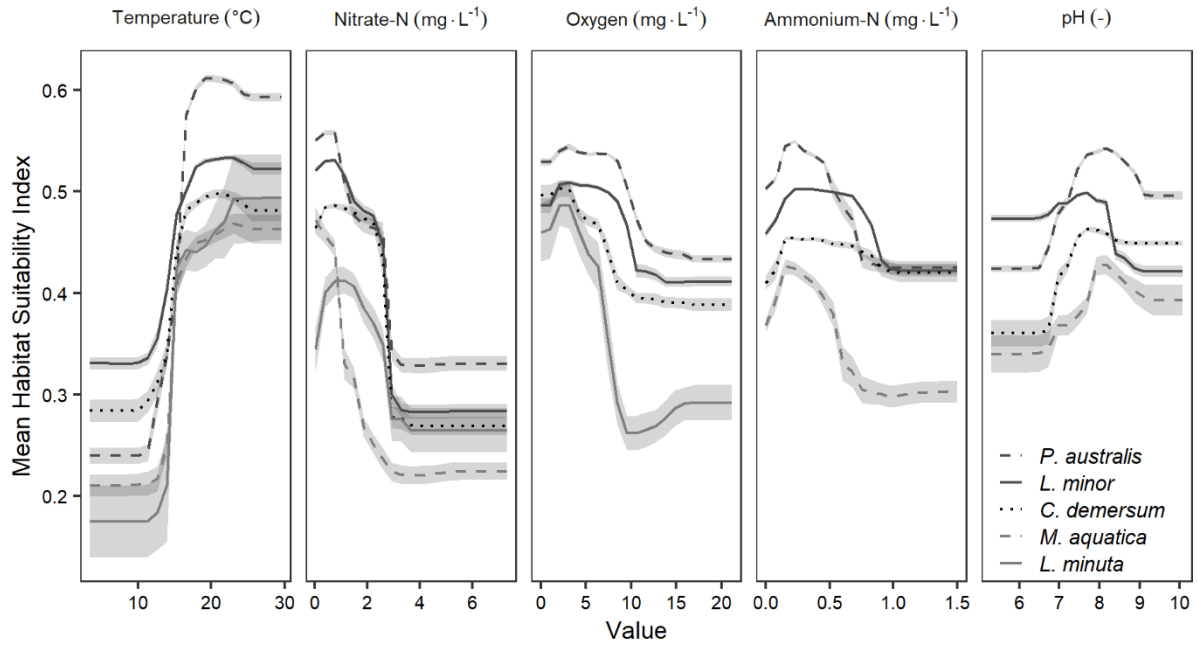

Figure 7.2: Partial dependence plots (PDPs) of the five most-informative variables for a selection of five macrophytes. *Plots were derived from macrophyte-specific optimised conditional random forests and show the inferred effect of an environmental variable on the habitat suitability for a specific species. An optimal range can be observed for each variable, with a general positive effect of temperature and negative effect of nitrogen. Some models did not contain all selected variables, resulting in an absence of a variable-specific influence plot.*

P. australis showed to be the most generalist species among the considered macrophyte species, often reflecting the highest average suitability score, except at low temperature and pH values (Figure 7.2). In contrast, *M. aquatica* frequently exhibited the lowest HSI scores, indicating a more specialist behaviour. *C. demersum* seemed to be the least affected, being partially consequential to the exclusion of chlorophyll *a*, sulphate and Kjeldahl-nitrogen (see Figure 7.1) throughout this analysis. Differences in habitat suitability scores between *L. minor* and *L. minuta* were generally higher at undisturbed conditions (i.e. low temperature, low nitrate and high oxygen concentrations) and tended to decrease towards higher disturbance (Figure 7.2), indicating a reduced discrepancy in habitat suitability due to nutrient pollution or overall climate change.

Similar partial dependence analyses were performed for all 58 macrophytes within the provided data set, though required the exclusion of one species as none of the selected variables were included in the developed model. The remaining 57 species showed similar patterns as observed for the selected subset, though averaging all species-specific responses caused relatively high deviation around the overall mean (Figure 7.3). This illustrates that preferences among macrophytes are similar regarding the main drivers and benefit from general guidelines, while additional fine-tuning is required when aiming for improving habitat suitability for a specific species.

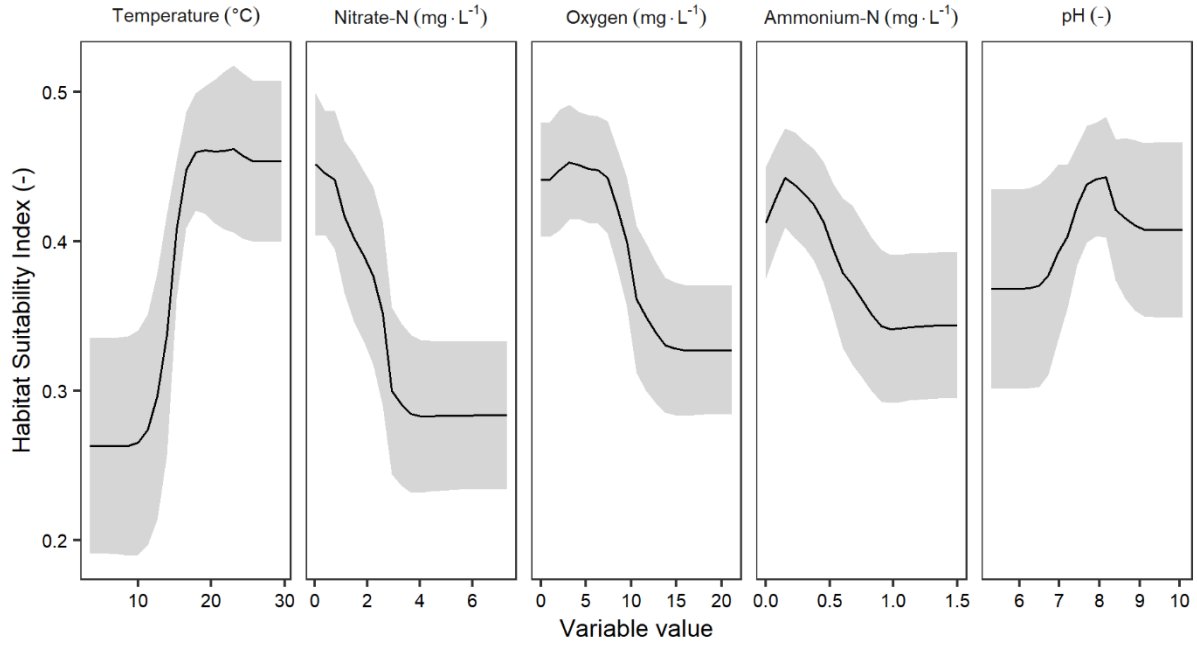

Figure 7.3: Partial dependence plots (PDPs) of the five most-influential variables for all macrophytes (N = 57). *The average influence of a specific environmental variable on habitat suitability (black line) follows a similar pattern as observed in Figure 7.2. Moreover, similar optimal ranges can be observed for each variable, with a general positive effect of temperature and negative effect of nitrogen. The grey ribbon depicts the standard deviation of the mean.*

7.3.3 Application of optimised models

Application of the optimised species-specific models on the complete data set suggested a suboptimal use of suitable habitats (Figure 7.4). Over time, an overall increase in suitable and occupied habitats was observed for each macrophyte, although the limited repeated temporal sampling clouds the presence of clear patterns (i.e. only a few sites were sampled more than once). Discrepancies between observations and predictions tended to increase with decreasing observed prevalence, showing a high degree of overlap for *P. australis* (period: 1990-2010) and a clear difference between the locations occupied by and available for *L. minuta* (Figure 7.4). No observations of *L. minuta* before 1999 were included in the common data, though the upward trend indicated a rising reporting frequency (Figure 7.4), which is likely to increase further as more locations will provide a suitable habitat and dispersal pressure rises. Temporal trends of all 58 macrophyte species show relatively similar patterns and can be found in Appendix, Figure D.5 and Figure D.6.

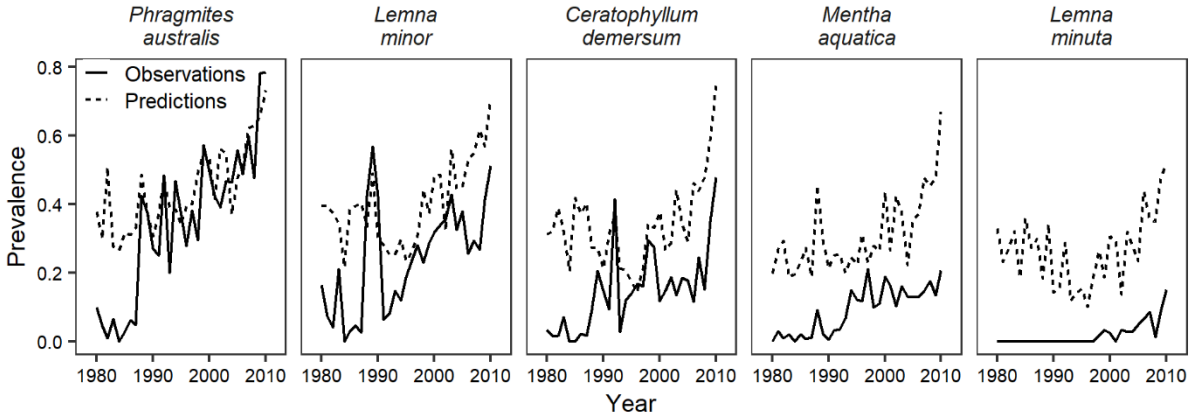

Figure 7.4: Temporal trend of observed and predicted prevalence of a selection of macrophytes. *Prevalence is determined by the fraction of sites where macrophyte presence is observed (solid line) or where conditions are suitable to support macrophyte presence (dashed line). The fraction of both suitable and occupied sites increases in time and indicates a suboptimal use of the available suitable habitats. Similar analyses of all 58 macrophyte species can be found in Appendix, Figure D.5 and Figure D.6.*

On average, abiotic conditions at the end of the sampling period (i.e. 2010) already supported relatively high habitat suitability scores (see Figure 7.4, Figure 7.5 and Appendix, Table D.2). The analyses suggested that, without any action being taken, suitability might commence dropping after 10 years (AVG-BAU), potentially due to inadequate nutrient concentrations. Indeed, when relying on a continuation of the temporal trend, nitrate concentrations dropped to 0 $mg{\cdot}L^{-1}$ (see Appendix, Figure D.4) and negatively influenced HSI (see Figure 7.2). In contrast, when management aimed at obtaining PDP-derived optimal conditions (see Figure 7.2), habitat suitability tended to remain relatively stable (AVG-KEY; Figure 7.5).

Polluted sites generally benefitted from any type of management, though indicated better absolute improvement in suitability with variable-specific action, especially with respect to *P. australis* and *M. aquatica* (EXT-KEY; Figure 7.5). Similarly, temporal analysis of the eutrophic systems suggested that a focus on managing key variables (NUT-KEY) provided higher habitat suitability scores compared to the business-as-usual (NUT-BAU) scenario (Figure 7.5).

Throughout these scenarios, highest suitability scores were generally observed for *P. australis*, while *M. aquatica* showed to be greatly affected by the prevailing nutrient conditions (Figure 7.5), thereby corroborating their relatively generalist and specialist behaviour, respectively. *C. demersum* was only limitedly affected by any type of management, except for the business-as-usual scenario towards average starting conditions (AVG-BAU; Figure 7.5), which is potentially linked with a different degree of dependence on the considered variables. *L. minor* and *L. minuta* showed relatively similar patterns regardless of the scenario, with generally higher suitability scores for *L. minor*, although comparable scores were observed when management focused on key variables under non-extreme starting conditions (AVG-KEY and NUT-KEY; Figure 7.5). Hence, a preference of both *Lemna* spp. towards the same abiotic conditions is to be expected.

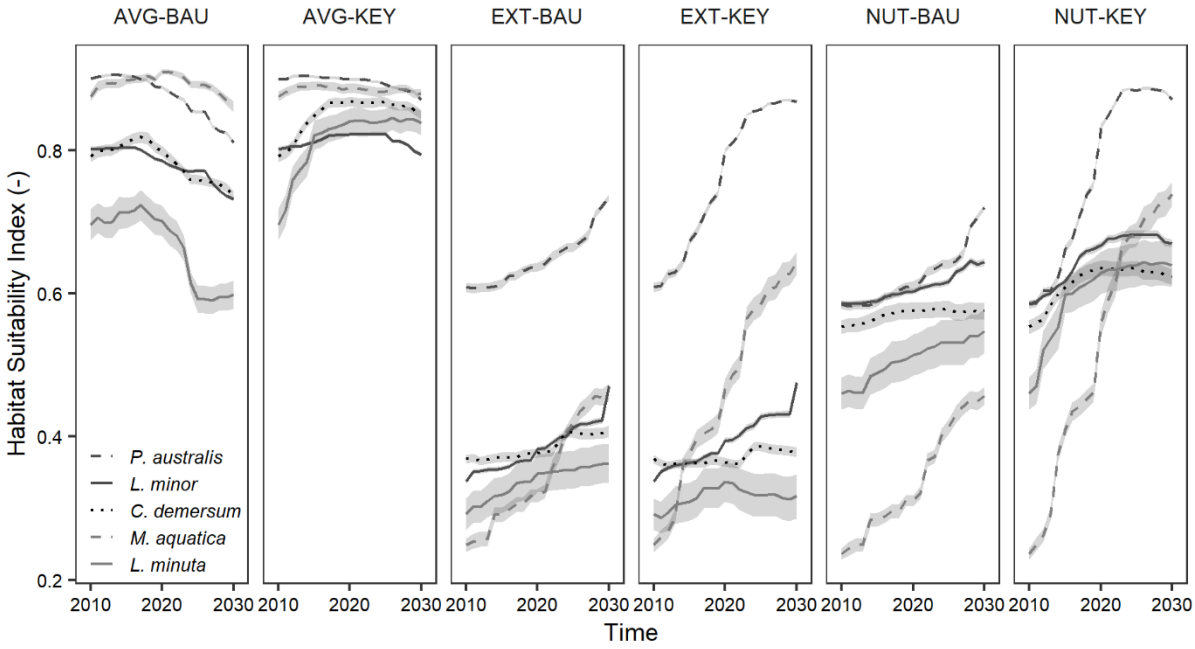

Figure 7.5: Effects of management and starting conditions on habitat suitability. *Management is generally beneficial, except for business-as-usual with average variable values. AVG: Average starting conditions; EXT: Extreme starting conditions; NUT: Nutrient-enriched starting conditions; BAU: Business-as-usual; KEY: Management focused on key variables (see Figure 7.2). To improve visualisation, standard errors (*N = *10) are depicted as grey ribbons instead of standard deviation.*

Similar to the partial dependence plots (Figure 7.2) and the management scenarios (Figure 7.5), higher suitability scores for *L. minor* occurred for the majority of locations (79.0 %) within the original data compared to *L. minuta*. However, not all sites with reported *L. minor* presence sustained lower HSI scores for *L. minuta* compared to *L. minor* and vice versa. About a quarter (28.3 %) of the locations with *L. minor* presence provided higher suitability scores for *L. minuta*, while even a higher fraction (39.0 %) of the sites occupied by *L. minuta* supported higher HSI scores for *L. minor* (Figure 7.6). The majority of sites (71.4 %) remained, however, unoccupied by either species, though showed generally higher HSI scores for *L. minor*. Moreover, the HSI frequency distribution of all unoccupied sites suggested that several sites provided suitable conditions for *Lemna* spp. presence, which additionally illustrates the suboptimal use of suitable habitats.

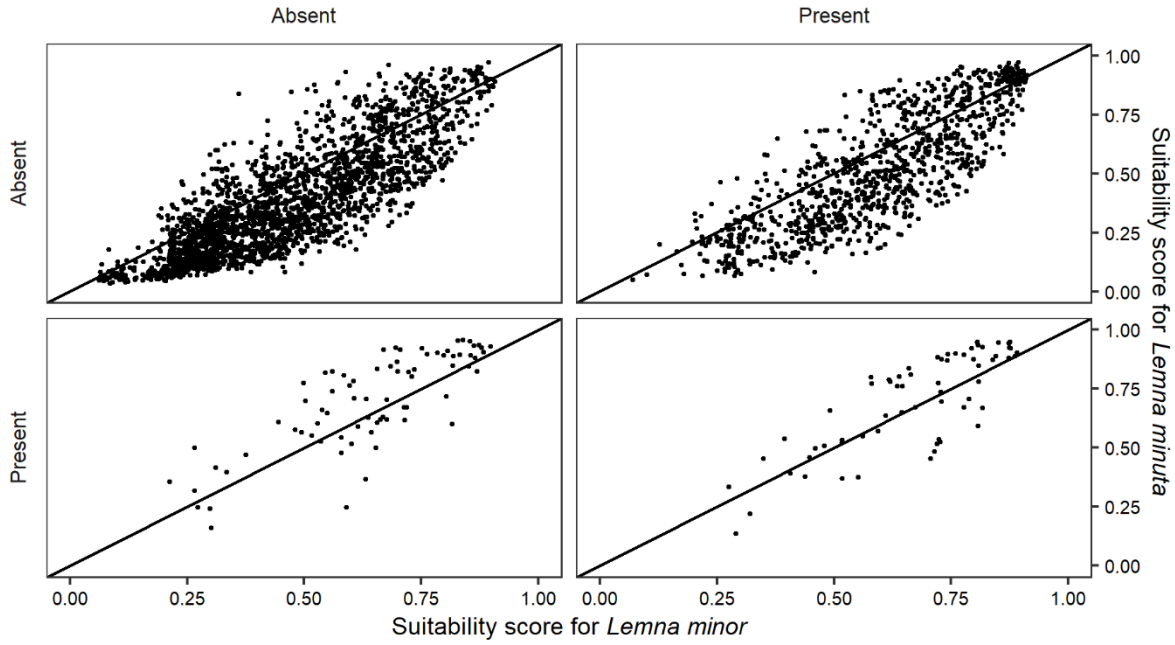

Figure 7.6: Habitat suitability of* Lemna minor *and* Lemna minuta *conditional to their occurrence. *Sites with absence of both* Lemna *spp. (top-left) cover a range of suitability scores and are mostly situated below the agreement line indicating that the majority of unoccupied sites provides slightly more suitable conditions for* L. minor. *Sites with observed* L. minor *presence and* L. minuta *absence (top-right) show a similar pattern, indicating slightly better conditions for* L. minor *and corroborate the observations. Sites with observed presence of* L. minuta *(bottom row) are situated on both sides of the agreement line and reflect similar suitability scores for both* Lemna *spp.*

7.4 Discussion

7.4.1 Model performance and variable importance

Overall, obtained models provided good discriminatory power and classification accuracy (Swets, 1988), while hyperparameter tuning hardly affected the selected performance indicators, suggesting that conditional random forests represent a valuable approach within ecological data-based modelling, even under default settings (see also Fox *et al.* (2017) and Freeman *et al.* (2015)). Higher performance scores for random forests have been reported in literature, though these tend to vary among applications (see Table 7.5).

Table 7.5: Comparison of the obtained AUC scores with reported literature. *Most studies rely on accuracy, Cohen's kappa or the True Skills Statistic (TSS) to complement AUC. [a]: mean value; [b]: median value.*

Topic	AUC	Reference
Spatial bird distributions in the USA	0.917[a] ± 0.076	Barbet-Massin *et al.* (2014)
Temporal bird distributions in the USA	0.896[a] ± 0.090	Barbet-Massin *et al.* (2014)
Fish distribution in lake ecosystems	0.891[b]	Guo *et al.* (2015)
Biotic interactions in fish distribution models	0.85 – 0.95	Vezza *et al.* (2015)
Distribution of European grayling	0.943[a] ± 0.005	Fukuda *et al.* (2013)

Still, model performance is potentially deflated due to the inclusion of false absences within both the training and test data. Such non-occupation of suitable habitats originates from a variety of ecological processes, including limited macrophyte dispersal and increased stochasticity of extinction due to spatial isolation (Demars and Edwards, 2009). The majority of these false absences were excluded during data pre-processing in order to reduce ambiguity and to avoid reduced model performance scores (Gallien *et al.*, 2012; Guisan and Theurillat, 2000). Yet, the lack of a clearly defined niche in combination with the trade-off between model performance and data loss impedes the elimination of all false absences. Hence, several suitable unoccupied sites remain in the training and test data, resulting in model misclassifications and reduced model performance. This deflation, on the other hand, is counteracted by the spatiotemporal autocorrelation of the test data (Araújo *et al.*, 2005a; Araújo *et al.*, 2005b; Elith and Leathwick, 2009), although the relative contribution of both biases remains unknown.

Importance-based variable ranking identified temperature as a major descriptor of habitat suitability, showing a positive effect on habitat suitability scores when increasing. This complies with literature reporting (i) temperature as best-predicting factor for macrophyte diversity (Demars and Edwards, 2009), (ii) growth limitation at low temperatures in clear lakes (Dale, 1986), (iii) an optimal range for photosynthetic activity between 20 °C and 35 °C (van der Heide *et al.*, 2006), (iv) higher invasion vulnerability at higher temperatures (Hussner, 2009) and (v) dense floating mats causing temperature increases (Netten *et al.*, 2010). Hence, an increase in temperature due to, for instance, climate change, can have a beneficial effect on macrophyte presence, although also negative effects due to soil anoxia and related stress have been observed (Genkai-Kato and Carpenter, 2005).

In contrast, suitability scores were negatively related with increasing nitrate (NO_3^-) and ammonium (NH_4^+) levels, reflecting the expected harmful effect of water pollution on macrophyte occurrence and diversity (Bakker *et al.*, 2013; Barker *et al.*, 2008; Scheffer *et al.*, 1993b). More specifically, under elevated nutrient levels, phytoplankton has the potential to grow rapidly and outcompete macrophytes by changing nutrient conditions and light penetration (Lu *et al.*, 2012; Scheffer *et al.*, 1993b).

Surprisingly, oxygen was selected among the five most informative variables to delineate the occupied abiotic habitat. Macrophytes are relatively independent of oxygen within the water column due to their inherent production capacity, though tend to reduce oxygen during nocturnal respiratory activity (Caraco and Cole, 2002; Carr *et al.*, 1997). Moreover, higher suitability scores were generally linked with reduced oxygen concentrations (i.e. around 4.5 $mg \cdot L^{-1}$), which often reflects reduced chemical water quality (Srebotnjak *et al.*, 2012). This observation is potentially caused by biotic feedback, which takes place when species occur in a specific environment and modify the prevailing abiotic conditions due to their presence (Vitousek *et al.*, 1997). For instance, the elevated HSI scores for the floating *L. minuta* at low-oxygen conditions might depict an effect of its presence on abiotic conditions (i.e. causing a drop in oxygen by limiting light penetration) rather than its presence being affected by low oxygen levels. Similarly, the presence of the floating alien *Eichhornia crassipes* negatively affected oxygen concentrations within the invaded tidal environment of the San Francisco Estuary (Tobias *et al.*, 2019), while the presence of the submerged alien *Elodea nuttallii* positively affected oxygen saturation within invaded lakes in Northern Ireland (Kelly *et al.*, 2015). Hence, the identified variable importance ranking merely reflects the capacity of the variable to delineate and describe the occupied habitats rather than providing information on steering behaviour. More specifically, no distinction can be made between variables that (1) affect macrophyte presence, (2) are affected by macrophyte presence and (3) combine both processes.

7.4.2 Temporal trends and future potential

Despite the annual fluctuations, positive temporal trends were observed for macrophyte prevalence within the study area. Both observed and predicted prevalence scores increased in time, while concentrations of the main pollutants (ammonium, nitrate, phosphorus) decreased (see Appendix, Figure D.3). This suggests that management efforts to reduce surface water pollution have provided positive results at the biotic level. However, these results should be interpreted with care as they are only valid under the assumption that sites were selected randomly (i.e. without any preference towards vegetated or non-vegetated sites). As this assumption might be too strict for specific years, it is considered likely that the depicted prevalence scores do not reflect the actual conditions, causing temporal patterns to fluctuate. More importantly, it is crucial to maintain management measures as (1) individual variables are often characterised by a wide range (see Appendix, Figure D.3) and (2) many surface waters in the Netherlands are still highly eutrophic (van Puijenbroek *et al.*, 2014).

Indeed, management measures positively influenced HSI scores for most macrophytes, especially when paying specific attention to altering the most descriptive variables (i.e. KEY management). A clear distinction with BAU management was observed in favour of KEY management, except when dealing with extremely polluted sites (EXT). This illustrated that the identification of key habitat descriptors can help in delineating management actions, but that case- and species-specific management actions are required for locations situated outside the realised niche. More importantly, it confirmed that macrophyte presence is influenced by a plethora of interacting variables (Bakker *et al.*, 2013; Demars and Edwards, 2009).

It should remain clear that the management scenarios in this study were composed by combining theoretical starting conditions and temporal patterns based on observed environmental conditions and patterns, respectively (see Section 7.2.2 and Appendix D.2). Hence, the resulting simulations merely illustrate the value of abiotic HSM towards scenario analysis and can be used to confirm and develop macrophyte-specific hypotheses. For instance, the high HSI scores for *P. australis* suggested a relatively high generalist behaviour, which has been illustrated by its highly invasive character (Bellavance and Brisson, 2010; Zedler and Kercher, 2004). Similarly, HSI scores for *M. aquatica* were strongly influenced by nutrient concentrations and suggested a more specialist behaviour, thereby contrasting reports on its presence in constructed treatment wetlands (Dhir *et al.*, 2009; Vymazal, 2013). Such characterisation is inherently nested in the study design, which resulted in the selection of generally occurring species (and, thus, the exclusion of actual specialist species from the study).

Throughout the simulated timeframe, HSI scores for the native *Lemna minor* and the alien *L. minuta* depicted relatively similar patterns and a decreased discrepancy when management focused on optimising the key descriptors, except for extremely polluted sites. This confirms field observations of both *Lemna* spp. coexisting and favouring similar environmental conditions (Ceschin *et al.*, 2016; Paolacci *et al.*, 2016), including a preference towards eutrophic conditions. However, due to the alien nature of *L. minuta*, it remains possible that the occupied environmental domain and associated model predictions underestimate the potential domain and habitat suitability scores (Gallien *et al.*, 2012). The upward temporal prevalence trends illustrate its endeavour to reach equilibrium and depict the so-called 'invasion debt' (Strayer, 2010). Moreover, simulations showed that pollution reduction supports increased habitat suitability for both *Lemna* species, implying a further increase in the future due to continuously decreasing nutrient concentrations (Blaas and Kroeze, 2016).

Both models and observations supported the coexistence of *L. minor* and *L. minuta* due to shared abiotic preferences. Yet, extrapolations to long-term natural conditions are to be performed with care as observations can be temporally biased and merely reflect a temporary situation. For instance, coexistence may also be caused by a disturbance-induced survival of *L. minuta* in a system dominated by *L. minor* or vice versa, thereby supporting temporary co-occurrence despite differences in species-specific habitat suitability. Such disturbances undermine the governing biotic resistance and increase the opportunity for natural succession, more diverse communities, higher productivity and nutrient retention, though simultaneously allow invasive (alien) species to establish (Demars and Edwards, 2009; Engelhardt and Ritchie, 2001; Strayer, 2010; Zedler and Kercher, 2005). Whether the observed co-occurrence of both *Lemna* spp. results in coexistence or outcompetition cannot be derived from the developed models and greatly depends on their autecological behaviour, functional traits and overall competitive strength (see also Figure 2.3) (Demars and Edwards, 2009; Kelly *et al.*, 2015; van Kleunen *et al.*, 2010).

Hence, more information from both controlled-conditions experiments and in-field observations is required to identify autecological behaviour and species interactions. For instance, functional traits like nutrient uptake rate and relative growth rate (RGR) can provide information on the invasive behaviour of a species (Njambuya *et al.*, 2011; van Kleunen *et al.*, 2010). Experiments performed on the invasive shrimp *Dikerogammarus villosus* and the native shrimp *Gammarus pulex* showed that the functional response (i.e. resource use) was higher for the invasive shrimp, thereby illustrating its observed invasive behaviour (Dodd *et al.*, 2014). The use of a similar index to infer invasive behaviour of alien macrophytes might prove useful within a proactive management framework.

7.4.3 Consequences for wetland and environmental management

Quantitative assessment of disturbances and macrophyte interactions and how these processes will change in the future remains a challenge when developing habitat suitability and species distribution models (Elith and Leathwick, 2009). Invasive alien species and climate change represent important threats to aquatic ecosystems, including freshwater wetland systems (Peterson *et al.*, 2008; Rahel and Olden, 2008; Walther *et al.*, 2009). For instance, dominance by invasive alien macrophytes has already caused the disappearance of native species due to light limitation, with additional negative effects on the macroinvertebrate community (Stiers *et al.*, 2011). Moreover, alterations in environmental conditions induced by climate change (e.g. increased temperatures, modified hydrological regimes) are expected to be advantageous towards invasive alien species and indicate an important interaction between two influential pressures (Rahel and Olden, 2008; Williams and Grosholz, 2008). In order to mitigate future impacts, it is imperative to develop contemporary wetland management plans that inhibit the establishment and spread of invasive species.

These management plans should encompass several focus points, including (1) the identification of locations with suitable abiotic conditions for non-invasive native species, (2) the identification of locations with suitable abiotic conditions for invasive species (both native and alien) and (3) the identification of species pools in the surrounding environment or within the sediment. The developed models in this study were able to identify key habitat descriptors and to infer overall habitat suitability conditional to the prevailing abiotic conditions for a selection of macrophyte species. For instance, abiotic habitat suitability for *Mentha aquatica* showed to be highly correlated with nutrient concentrations (nitrate, ammonium and phosphorus), while its prevalence increased in time due to a reduction in nutrient levels (see Figure 7.1, Figure 7.4 and Figure D.3). Hence, additional nutrient reduction within eutrophic treatment wetland benefits habitat suitability for *M. aquatica*.

Similarly, habitat suitability scores for the submerged *Ceratophyllum demersum* showed to be less affected by the main habitat descriptors, when compared to the other selected macrophytes. This reduced relation is potentially caused by the exclusion of chlorophyll *a*, sulphate and Kjeldahl-nitrogen from the dependency analysis and suggests that *C. demersum* is less sensitive towards generic alterations of the abiotic conditions (i.e. focusing on the key habitat descriptors as depicted in Figure 7.5). Hence, a more species-specific analysis and management is needed to significantly affect habitat suitability for *C. demersum* and the associated chance of establishment (conditional to its introduction).

In addition to a local assessment of the available and required abiotic conditions, awareness on the presence of a local species pool is essential to decide between natural succession or manual introduction in order to obtain augmented species richness. Limited dispersal and connectivity have affected various restoration projects that relied on seed banks within the sediment or the proximity of local species pools to commence colonisation after abiotic restoration (Bakker *et al.*, 2013; Hilt *et al.*, 2006). Both processes support natural biotic restoration, though are often beneficial for highly-competitive generalist species, causing communities with low diversity and high biomass (Engelhardt and Ritchie, 2001). In absence of a viable seed bank, introduction greatly depends on the available direct (e.g. connected water bodies) or indirect (e.g. wind- or animal-induced) dispersal pathways (Murphy *et al.*, 2019).

However, only a fraction of the introduced propagules survives the prevailing abiotic conditions, being subsequently exposed to biotic interactions, including herbivory and (if present) the already established macrophyte community (Bakker *et al.*, 2013; Levine *et al.*, 2004), being conceptually visualised in Figure 7.7. Manual introduction can be considered when both abiotic and biotic conditions support the species' presence, though requires prior investigation on the reasons of their current absence (Bakker *et al.*, 2013; Bornette and Puijalon, 2011). For instance, high herbivory pressure in lakes or wetlands causes macrophytes to be absent and renders many re-stocking actions into failure when the pressure remains unaccounted for (Körner and Dugdale, 2003). Similarly, highly turbid water conditions caused by sediment-disturbing fish and crustaceans provide a poor basis for artificial introduction (Hilt *et al.*, 2006; Strayer, 2010). Hence, despite providing suitable abiotic conditions, the probability of successful natural succession can be low due to dispersal limitation and biotic interactions (see Figure 7.7).

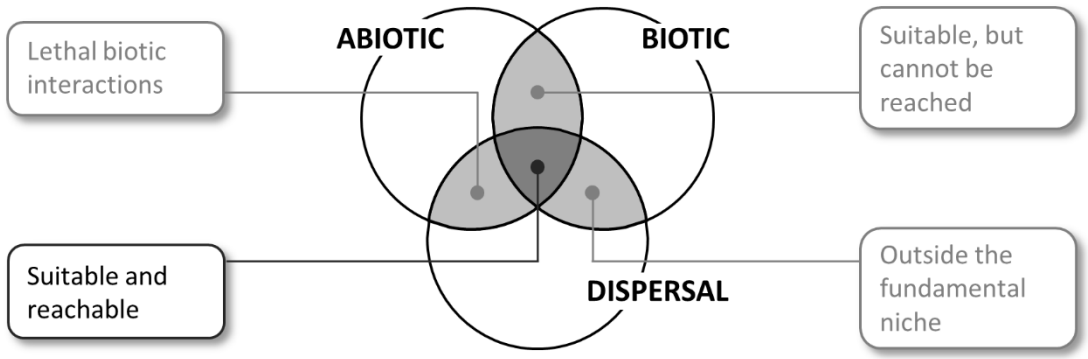

Figure 7.7: Conceptual visualisation of the contributing factors underlying macrophyte presence. *Both abiotic and biotic conditions need to be suitable for a species to occur, but they also need to be reachable to allow natural introduction. Manual introduction avoids the restriction implied by dispersal and thereby creates more options.*

7.4.4 Contribution to the study objective

The aim of this chapter was to combine ecological restoration and invasive alien species management by defining the effect of species-specific key variables on habitat suitability and elaborating on management options to optimise abiotic conditions. By means of correlative models, macrophyte occurrence data within the Limnodata Neerlandica were linked with the prevailing abiotic conditions in order to infer species-specific descriptions of the preferred habitats. These results help identifying species that possess the potential to thrive in the physicochemical conditions that are present within the considered wetland (see Section 1.2.1). Moreover, they illustrate how abiotic conditions can be changed to improve the habitat suitability for a specific macrophyte species, which additionally allowed the assessment of temporal trends and management scenarios on the habitat suitability of both native and alien species.

Variable-specific effects on habitat suitability often remained below HSI scores of 0.55 (see Figure 7.2), indicating that a single variable can create relatively unsuitable conditions and confirming that a concert of variables is needed to provide a suitable habitat (Bornette and Puijalon, 2011; Demars and Edwards, 2009). Hence, a holistic approach that targets a range of variables (e.g. wastewater treatment to reduce organic pollution, buffer strips in agricultural area to reduce nutrient input) to reduce pollutant concentrations positively affects habitat suitability for macrophytes. Increased macrophyte occupancy over time supports these inferences and highlights the positive impact of improved water management on macrophyte presence. Yet, the discrepancies between the observed and predicted prevalence suggest a temporal lag between abiotic restoration and biotic colonisation, which has also been observed in several other restoration projects (Bakker *et al.*, 2013; Jähnig *et al.*, 2011; Verdonschot *et al.*, 2013).

The models that were developed in this chapter allowed to infer (1) the most influential descriptors to delineate the occupied habitats, (2) the values of these key descriptors to provide optimal habitat suitability and (3) the effect of different management scenarios on species-specific habitat suitability scores. Based on these results, the value of data-driven modelling towards supporting freshwater management is illustrated. Moreover, within the defined study objective (see Section 1.2.1), nutrient conditions are assumed to be elevated and thereby resemble the starting conditions of the NUT scenarios. As temporal improvements in these scenarios support increased habitat suitability, a similar effect can be expected along the flow path through a constructed treatment wetland. This is especially interesting towards the implementation of zonation within the wetland, though remains threatened by competitive generalist species that have a tendency to create dense monocultures (e.g. *Phragmites australis*). By combining these models and field assessments of local species pools, a list of potential harmful or unwanted species (both native and alien) can be composed.

7.5 Conclusion

Conditional random forests (CRFs) showed to be a valuable approach for determining first-level habitat suitability scores, providing good performance and significantly outperforming null models while performance improvement via hyperparameter optimisation remained limited. Importance-based variable ranking differed between macrophytes, with temperature and nitrate as recurring key variables among the selected species. Nevertheless, a holistic approach tackling multiple variables at once is requested to obtain a significant increase in habitat suitability as the effect of a single variable remains relatively small. Further improvements of the developed abiotic habitat suitability models require laboratory tests and extensions with biotic information including nutrient use, biomass production, dispersal dynamics and potential allelopathic behaviour. This need was illustrated by the observation that some sites were characterised by higher suitability scores for *L. minuta* while *L. minor* was observed and vice versa.

8

Functional response and relative growth rate to assess invasiveness[6]

Highlights

- Functional response is insufficient to forecast invasive behaviour
- Relative growth rate was similar among both *Lemna* species
- Low nutrient requirement and high fresh weight indicate invasiveness

[6] This chapter is based on Van Echelpoel, W.; Boets, P. and Goethals, P. L. M. (2016) Functional response (FR) and relative growth rate (RGR) do not show the known invasiveness of Lemna minuta (Kunth). *PLoS ONE* **11**, e0166132, doi: 10.1371/journal.pone.0166132.

Abstract

Growing travel and trade threatens biodiversity as it increases the rate of biological invasions globally, either by accidental or intentional introduction. Therefore, avoiding these impacts by forecasting invasions and impeding further spread is of utmost importance. In this study, three forecasting approaches were tested and combined to predict the invasive behaviour of the alien macrophyte *Lemna minuta* in comparison with the native *Lemna minor*: the functional response (FR) and relative growth rate (RGR), supplemented with a combined biomass-based nutrient removal (BBNR). Based on the idea that widespread invasive alien species are more successful competitors than native species, a higher FR and RGR were expected for the alien compared to the native species. Five different nutrient concentrations were tested along a nitrogen (4 $mg{\cdot}L^{-1}$ up to 70 $mg{\cdot}L^{-1}$) and phosphorus (1 $mg{\cdot}L^{-1}$ up to 21 $mg{\cdot}L^{-1}$) gradient. After four days, a significant amount of nutrients was removed by both *Lemna* spp., though significant differences among *L. minor* and *L. minuta* were only observed at lower nutrient concentrations (i.e. lower than 17 $mg{\cdot}L^{-1}$ for nitrogen and 6 $mg{\cdot}L^{-1}$ for phosphorus) with higher nutrient removal exerted by *L. minor*. The derived FR did not show a clear dominance of the invasive *L. minuta*, contradicting field observations. Similarly, the RGR ranged from 0.4 d^{-1} to 0.6 d^{-1}, but did not show a biomass-based dominance of *L. minuta* (i.e. 0.5 ± 0.3 d^{-1} versus 0.6 ± 0.2 d^{-1} for *L. minor*). BBNR showed similar results as the FR. Contrary to the expectations, all three approaches resulted in higher values for *L. minor*. Consequently, based on our results FR is sensitive to differences, though contradicted the expectations, while RGR and BBNR do not provide sufficient power to differentiate between a native and an invasive alien macrophyte and should be supplemented with additional ecosystem-based experiments to determine the invasion impact.

8.1 Setting the scene

In Chapter 7, the potential threat of *Lemna minuta* towards ecosystem conservation has been suggested by the decreased discrepancy in habitat suitability index when disturbance increases (i.e. higher temperatures and nitrate concentrations). However, these inferences are highly dependent on occurrence data within the invaded range, which often violate the equilibrium assumption and underestimate the species' realised niche (Gallien *et al.*, 2012; Guisan and Zimmerman, 2000). Alternative approaches consider the implementation of pre-introduction procedures and the study of species-specific traits, which require a completely different setup, but are crucial to counter current introduction rates.

Identifying potential introductions, avoiding establishment and impeding further spread of invasive alien species (IAS) by detection and subsequent large-scale eradication requires commitment, financial input and highly destructive measures (Myers *et al.*, 2000). As not all traits of the invader are known, new functions can be introduced without changing the community composition drastically (e.g. niche differentiation resulting in an increase in total ecosystem biomass) (Vilà and Weiner, 2004). However, this introduction of completely new traits is limited (Funk and Vitousek, 2007), underlining that knowledge and early detection is required from a conservation point of view.

Forecasting invasion impact is a challenge in invasion biology (Dick *et al.*, 2013; Levine *et al.*, 2003; Pyšek and Richardson, 2007), as each organism interacts differently with its surrounding (Vitousek *et al.*, 1997), making it hard to determine a general effect of biological invasions. With enhanced competition being theorised as a major mechanism supporting successful invasion (Levine *et al.*, 2003), several authors have been investigating the competitive interaction between native and alien species as a first sign of alien or native dominance (e.g. Vilà and Weiner (2004), Njambuya *et al.* (2011), Gioria and Osborne (2014)).

Such a competitive advantage depends on a difference in functional identity, which is hypothesised to be involved in determining the final impact of invasion (Gooden and French, 2015; Levine *et al.*, 2003). Successful invasions generally occur when the non-native species displays higher values for competitively advantageous traits, while the intensity of the advantage is defined by the difference between the trait values. Therefore, approaches describing a difference in one (or more) functional trait(s) are applied to predict a species' invasive behaviour, for instance the functional response (FR), relative growth rate (RGR), nutrient content and specific leaf area (SLA) (Dick *et al.*, 2013; Gioria and Osborne, 2014; Grotkopp *et al.*, 2002; Pyšek and Richardson, 2007).

These differences in functional traits are also expected to be expressed at the sub-individual level (e.g. cellular, molecular, histological), for which (sub-)cellular biomarkers can be used to identify the factors that influence invasive behaviour (Colin *et al.*, 2016). Such biomarkers allow to measure and evaluate changes at the cellular, biochemical or molecular level in response to specific external signals (e.g. environmental conditions) (Mayeux, 2004). Despite being able to identify changes at the sub-individual level, the appropriate extrapolation of these biomarker-based results to the population and community level remains unclear (Friberg *et al.*, 2011). Moreover, considering a high physiological linkage, a similar response among different species is to be expected and can challenge the observation of significant differences (Colin *et al.*, 2016). An additional drawback of this technique is the poor knowledge of appropriate biomarkers for investigating macrophyte species (Brain and Cedergreen, 2008). Therefore, subsequent selection of the FR and RGR is based on their applicability, their ease of application, their link with population and community dynamics, and their focus on either input (resource use, FR) or output (biomass production, RGR).

The functional response is a known concept in general ecology, but it is only recently introduced in invasion ecology for comparing the per-capita resource uptake rate of native and alien species in function of the resource density (e.g. Alexander *et al.* (2014), Dick *et al.* (2013), Haddaway *et al.* (2012) and Médoc *et al.* (2015)). It states that an invasive alien species has a higher functional response compared to the native, because of its higher resource use efficiency (Dick *et al.*, 2013). In contrast to the functional response, which focuses on resource use (input-based), the relative growth rate focuses on the increase in biomass (output-based) to determine the invasion potential of an alien species and is considered as a proxy for the species' fitness (Gioria and Osborne, 2014). Therefore, several authors have been investigating the difference in RGR between native and alien species to predict the invasion potential of an alien species (e.g. Gérard and Triest (2014), Njambuya *et al.* (2011), Riley and Dybdahl (2015)). Application of the RGR to determine the invasive potential of macrophytes is rather limited to rooted macrophytes (e.g. Barrat-Segretain (2005), Eller *et al.* (2015), Hussner (2009)), with less attention towards floating macrophytes (e.g. Netten *et al.* (2010), Njambuya *et al.* (2011)). In contrast, the implementation of the FR concept is rare with respect to macrophyte assessment, though has proven to be successful for fish and macroinvertebrates (e.g. Alexander *et al.* (2014), Dodd *et al.* (2014)).

Within this chapter, attention is given to resource- and output-based macrophyte traits to infer their applicability for forecasting the invasive behaviour of an alien species. The aim is to determine species-specific results for the functional response and relative growth rate and to establish result similarity. By tackling these issues, an answer is provided to RQ3.1, as defined in Chapter 1. Hence, this chapter concludes with a statement on the applicability of the selected traits.

8.2 Materials and methods

8.2.1 Experimental setup

A pure culture of *L. minor* was ordered from Blades Biological (United Kingdom, http://www.blades-bio.co.uk). *L. minuta* was collected from the Bourgoyen nature reserve (51.062253, 3.673827), situated near Ghent (Belgium). About 20 fronds of each species were placed separately in a nutrient medium based on OECD and ISO guidelines for chemical testing with *L. minor* and is referred to as the full strength modified Steinberg medium (OECD, 2006). Fluorescence lamps provided 16 hours of light, followed by 8 hours of darkness, with an intensity of 45 $\mu mol \cdot m^{-2} \cdot s^{-1}$ up to 58 $\mu mol \cdot m^{-2} \cdot s^{-1}$. Temperatures of the growth medium varied between 21.6 °C and 24.0 °C. Every two to three days, new medium was provided and aquaria were rinsed thoroughly with tap water. Fronds showing the start of algae growth were removed or rinsed carefully. Selected *Lemna* spp. plants were grown in these conditions for at least two weeks to acclimate.

The tests were performed with similar light and temperature conditions as the aforementioned growth conditions. All recipients were covered at the sides with aluminium foil to constrain algae growth. The original modified Steinberg medium (C_0) was diluted with deionised water to obtain the following series of concentrations: C_0, $0.5 \cdot C_0$, $0.25 \cdot C_0$, $0.125 \cdot C_0$, and $0.0625 \cdot C_0$, hereafter referred to as: C1, C2, C3, C4, and C5, respectively. The composition of the growth medium within these concentration classes is described in Table 8.1.

Table 8.1: Composition and gradient of the growth medium used for performing the experiment. *The composition of C1 is based on the Steinberg medium used for chemical testing with* Lemna minor *(OECD, 2006).*

	C1	C2	C3	C4	C5
Macronutrients ($mg \cdot L^{-1}$)					
KNO_3	350	175	87.5	43.75	21.875
KH_2PO_4	30	15	7.5	3.75	1.875
K_2HPO_4	4.2	2.1	1.05	0.525	0.2625
$MgSO_4$	49	24.5	12.25	6.125	3.0625
$Ca(NO_3)_2$	205	102.5	51.25	25.625	12.8125
Micronutrients ($\mu g \cdot L^{-1}$)					
H_3BO_3	120	60	30	15	7.5
$ZnSO_4$	100	50	25	12.5	6.25
Na_2MoO_4	40	20	10	5	2.5
$MnCl_2$	130	65	32.5	16.25	8.125
$FeCl_3$	456	228	114	57	28.5
Na-EDTA	1500	750	375	187.5	93.75

Of each concentration, 0.25 L was poured into a glass recipient and about 500 mg fresh weight of *L. minor* or *L. minuta* was added, along with a control series without vegetation. Determination of the fresh weight was performed by collecting biomass on a sieve and blotting the fronds with tissue paper to extract attached water as much as possible. Each test lasted for four days (96 h), based on a preliminary assessment, and was performed in triplicate, resulting in a total of 45 recipients per test. In total, two tests were run, resulting in six replicates for each treatment and a total of 270 measurements. A schematic overview of the experimental set-up for a single series is depicted in Figure 8.1.

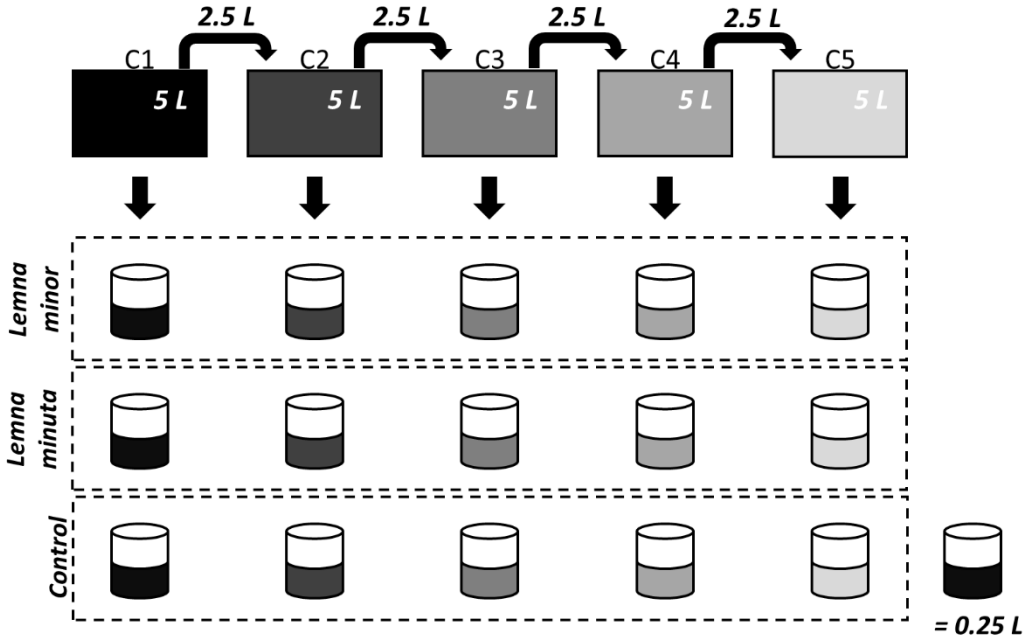

Figure 8.1: Schematic overview of the experimental set-up. *Relative initial nutrient concentrations are shown on top and were sampled at the start. The darkness within the aquariums represents the dilution state of the growth medium (black equals original modified Steinberg medium). Each recipient was filled with 0.25 L of its respective nutrient concentration and was performed in triplicate.*

8.2.2 Data collection

Growth medium samples were collected at the beginning and at the end of the test and stored at 4 °C in the dark prior to analysis. Within 36 hours after sampling, nutrient analysis was performed spectrophotometrically using Merck field kits for total nitrogen (test kits 1.14963.0001 and 1.14773.0001, operational range from 0.5 to 20 $mg \cdot L^{-1}$) and total phosphorus (test kit 1.14541.0001, operational range from 0.05 to 5 $mg \cdot L^{-1}$). For each batch, a blank and standard were used to determine the background signal and overall test efficiency, respectively. Medium samples of C1, C2, and C3 were diluted ten times with deionised water to comply with the operational range of the test kits. For each sample, the average of three measurements was used for further calculations.

Initial dry weight content was determined by drying representative subsamples of both *L. minor* and *L. minuta* for at least 48 hours at 60 °C (OECD, 2006). After two days, plant total fresh weight was determined and adapted to about 500 mg in each sample, as to keep biomass as constant as possible (FR is considered as the per-capita resource uptake). Leftover biomass was weighed and dried (48 hours at 60 °C) to determine the dry weight content and the estimated overall dry weight after two days of exposure. After four days, *Lemna* plants were harvested to determine both fresh weight and dry weight (48 hours at 60 °C).

8.2.3 Calculating characteristic values

Based on the obtained nutrient concentrations, nitrogen and phosphorus mass (expressed as mg) were derived by taking into account the volume of growth medium (0.25 L). Absolute nutrient removal was determined as the difference in initial and final nutrient mass. For this, the initial nutrient mass was determined as the average of all six replicates per concentration, as each replicate originated from the same batch of (diluted) growth medium. Finally, the functional response (nutrient mass removed in function of initial nutrient concentration) was determined. Next to the absolute nutrient removal, relative nutrient removal (*RNR*) was calculated (Equation 8.1) for each individual sample.

$$RNR = \frac{(m_{0,avg} - m_4)}{m_{0,avg}} \cdot 100\% \qquad \text{(Equation 8.1)}$$

With *RNR* the relative nutrient removal (%), $m_{0,avg}$ the average nutrient mass at day 0 (mg) and m_4 the nutrient mass at day 4 (mg).

The (estimated) dry biomass after exposure was determined after two and four days and compared with the initial (at day 0) and adapted (at day 2) dry weights, respectively. Similar to the observed nutrient removal, biomass increase was expressed both in absolute (dry weight increase) and relative (relative growth rate) terms of which the latter was calculated based on Equation 8.2, representing the relative growth rate (RGR) between day 2 and day 4.

$$RGR = \frac{\ln DW_4 - \ln DW_2}{t} \qquad \text{(Equation 8.2)}$$

With *RGR* the relative growth rate (d^{-1}), DW_4 the dry weight after four days (mg), DW_2 the adapted dry weight after two days (mg) and t the time interval (d).

Subsequently, nutrient removal and biomass increase were combined in a single variable to determine a more species-specific nutrient removal. Nutrient removal was expressed per gram biomass, with the latter being rather dynamic, resulting in three different values: initial dry weight, final dry weight and net dry weight increase.

The net dry weight increase was used under the assumption that duckweed allocates nutrients directly for new biomass instead of enriching already existing biomass (Körner and Vermaat, 1998). This suggests that an increase in nutrient uptake is directly related to an increase in biomass production. Follow-up of this nutrient uptake per gram newly created biomass allows to determine whether new biomass has a continuous nutrient content or whether additional nutrients are stored. A species with a higher storage capacity has an advantage towards future disturbances. To determine this biomass-based nutrient removal (BBNR), Equation 8.3 was applied.

$$BBNR = \frac{m_{0,avg} - m_4}{(DW_4 - DW_{2,ad}) + (DW_2 - DW_0)} \quad \text{(Equation 8.3)}$$

With *BBNR* the biomass-based nutrient removal expressing nutrient mass removed per unit biomass ($mg \cdot g^{-1}$), $m_{0,avg}$ the average initial nutrient mass (mg), m_4 the final nutrient mass (mg), DW_4 the biomass dry weight after four days (g), $DW_{2,ad}$ the estimated biomass dry weight at the beginning of the second period of two days (g), DW_2 the estimated biomass dry weight at the end of the first two days (g) and DW_0 the estimated initial biomass dry weight (g).

8.2.4 Statistical analysis

Obtained data of both tests were merged into a single data set and subsequently analysed using MS® Excel® and RStudio (R Core Team, 2016; RStudio Team, 2015). Outliers were identified by Cleveland dotplots and boxplot construction (Zuur *et al.*, 2010), though were initially not removed from the data set prior to subsequent statistical analysis. Not removing any value from the data set was based on the fact that all analyses were performed by the author and that spatial randomisation was applied when possible, thereby limiting the amount of valid arguments for outlier removal. During a second run, extreme values were removed to investigate their influence on the reported results.

Secondly, normality was tested using the Shapiro-Wilk test. When no significant difference from the normal distribution was observed ($p > 0.05$), paired *t*-tests were performed, in all other cases ($p < 0.05$) the paired Wilcoxon signed-rank test was applied. All *p*-values were considered as part of a multiple comparison set-up, for which a correction of the significant threshold value (α) is required. This correction is necessary as multiple comparisons increase the odds of observing a significant difference, though it increases the possibility of a type II error (accepting the null hypothesis while the alternative hypothesis is correct) (Armstrong, 2014). In short, a Bonferroni correction was applied for determining a new threshold value for each batch of five comparisons (i.e. $\alpha = 0.01$).

8.3 Results

8.3.1 Nutrient removal

Nutrient analyses performed at day 0 and day 4 resulted in the average nutrient concentrations provided in Appendix (Table E.1 and Table E.2) for total nitrogen (tN) and total phosphorus (tP), respectively. Recovery of a standard solution ranged from 93 to 99 % for nitrogen and from 95 to 98 % for phosphorus. As the initial nitrogen concentration of C5 (i.e. 4.2 ± 0.1 $mg \cdot L^{-1}$) was already quite low, measurements of the final nitrogen concentrations happened to be below the detection limit of 0.5 $mg \cdot L^{-1}$. These results were set to zero prior to determining average nitrogen concentration.

Subsequently, nitrogen and phosphorus mass (expressed in mg) were inferred from the measured nutrient concentrations (volume of 0.25 L), resulting in a similar nutrient content for *L. minor* and *L. minuta* (see Figure 8.2 and Figure 8.3). Both total nitrogen and total phosphorus differed significantly (*p*-values < 0.01) from the initial mass when *L. minor* or *L. minuta* was present at high (concentration C1) or low (concentration C5) nutrient concentrations (see Table 8.2). At intermediate concentrations, both significant and non-significant differences were observed (see Table 8.2).

The reference series (i.e. no plants) did not show a significant difference (all *p*-values > 0.01) for nitrogen mass, though some series (C1 and C2) showed a significant difference (*p*-values < 0.01) for phosphorus mass. Correcting for the analysis efficiency (based on the recovery of a standard solution), however, resulted in *p*-values not exceeding the threshold level of 0.01. Consequently, it can be stated that, in general, the presence of both *Lemna minor* and *Lemna minuta* significantly affected the nutrient content of the provided growth medium.

Table 8.2: Obtained* p*-values after comparing initial and final nutrient masses. *Significant differences (*p < *0.01) can be found at high (C1) and low (C5) nutrient concentrations and at several intermediate nutrient concentrations.*

	Nitrogen		**Phosphorus**	
	L. minor	***L. minuta***	***L. minor***	***L. minuta***
C1	< 0.001	< 0.001	0.002	< 0.001
C2	0.031	0.031	0.001	0.031
C3	0.31	0.007	< 0.001	< 0.001
C4	< 0.001	< 0.001	0.031	< 0.001
C5	< 0.001	< 0.001	< 0.001	< 0.001

No significant differences in nutrient removal were found between *L. minor* and *L. minuta*, except for nitrogen at concentration C4 (t = -5.3557, df = 5, p = 0.003) and phosphorus at concentration C3 (t = -6.1281, df = 5, p = 0.002) (see Figure 8.2, Figure 8.3 and Table 8.3). Relative nutrient removal, as calculated with Equation 8.1, showed that at low concentrations, relatively more nutrients were removed (Figure 8.4). Still, a slightly higher relative removal was observed for *L. minor* in comparison with *L. minuta*, with similar significant differences for nitrogen at concentrations C4 and for phosphorus at concentration C3. In short, the FR is able to identify a difference in nutrient removal, though it is limited to only one out of five concentration levels for each nutrient.

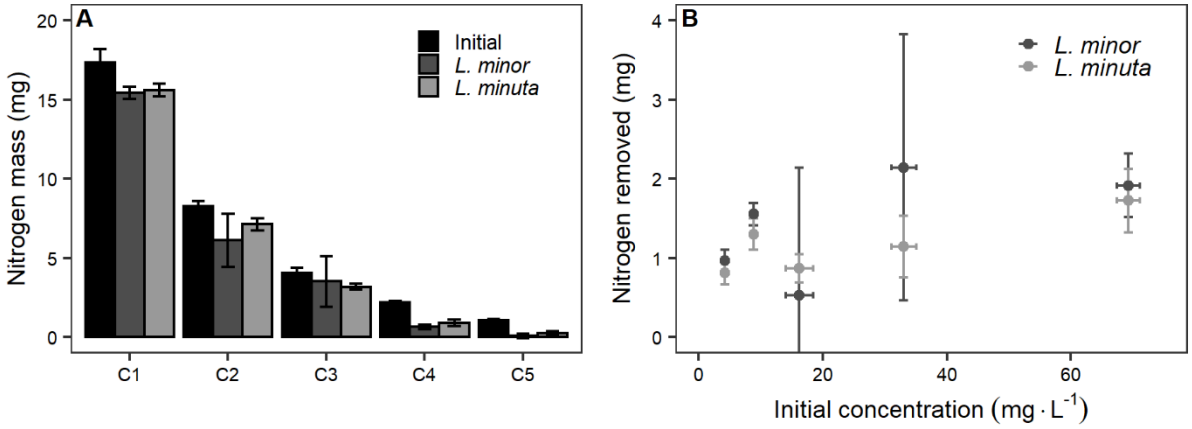

Figure 8.2: Absolute nitrogen removal by* L. minor *(dark grey) and* L. minuta *(light grey). *A: Nitrogen mass present at beginning (black bars) and after four days (grey bars). B: Amount of nitrogen removed in function of the initial amount of nitrogen, representing the functional response.*

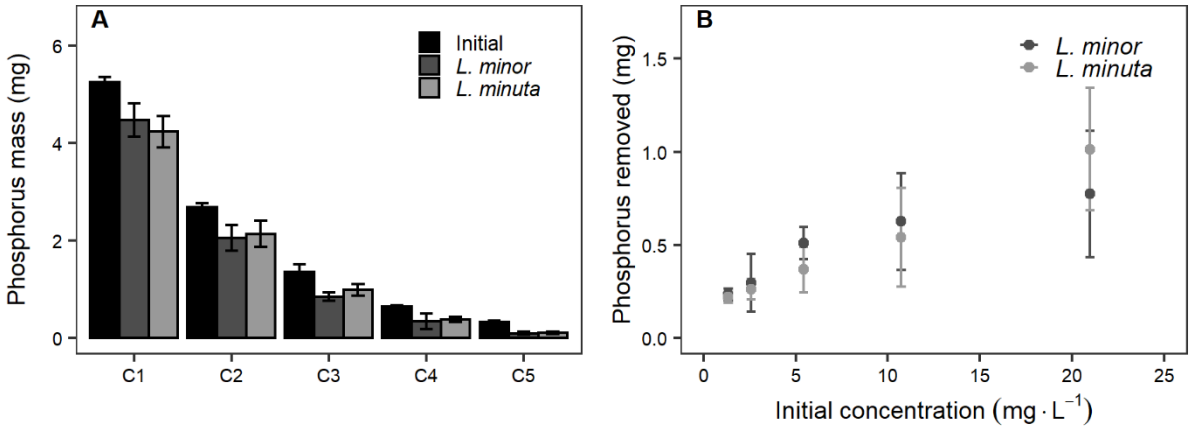

Figure 8.3: Absolute phosphorus removal by* L. minor *(dark grey) and* L. minuta *(light grey). *A: phosphorus mass present at beginning (black bars) and after four days (grey bars). B: amount of phosphorus removed in function of the initial amount of phosphorus, representing the functional response.*

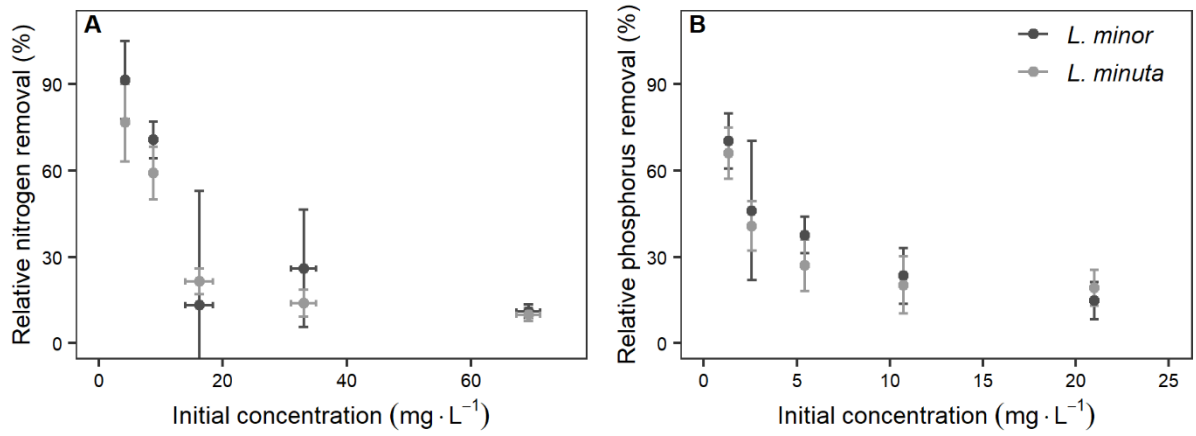

Figure 8.4: Relative removal of nutrients by* L. minor *(dark grey circles) and* L. minuta *(light grey circles). *A: nitrogen removal. B: phosphorus removal. At low nutrient concentrations relatively high nutrient removal efficiencies are observed.*

8.3.2 Biomass increase

At three different moments in time (day 0, day 2 and day 4) both fresh and dry weight of *Lemna* biomass were determined, with biomass dry weight at day 0 and day 2 being estimations based on the observed dry matter content of collected subsamples. Six samples (three for each species) were removed from the dataset as not enough biomass was present to determine the dry weight content. The resulting average dry weights (estimations, except for day 4) are provided in Appendix (Table E.3 and Table E.4).

The increase in biomass dry weight of *L. minor* between day 2 and day 4 was relatively similar among different concentrations (all *p*-values > 0.01) as it ranged from 30 ± 10 mg at concentration C4 to 35 ± 4 mg at concentration C1. In contrast, there was more fluctuation in the biomass increase of *L. minuta*, showing the highest increase in dry weight (32 ± 7 mg) at concentration C2 and the lowest increase (18 ± 8 mg) at concentration C4 (see Figure 8.5), though no significant difference was observed.

These fluctuations became less severe when considering the relative growth rate of *L. minuta*, ranging from 0.4 ± 0.2 d^{-1} at concentration C4 to 0.5 ± 0.3 d^{-1} at concentration C5 without any significant difference (all *p*-values > 0.01). In contrast, the relative growth rate of *L. minor* fluctuated more when compared with its related absolute biomass increase, as it ranged from 0.5 ± 0.1 d^{-1} at concentration C3 to 0.6 ± 0.2 d^{-1} at concentrations C1 and C5 (see Figure 8.5). Nevertheless, these growth rates were considered to be similar as no significant difference was observed (all *p*-values > 0.01).

Net biomass increase between day 2 and day 4 differed significantly between *L.* minor and *L. minuta* at concentration C4 (t = 5.3484, df = 4, p = 0.006) (Figure 8.5). In contrast, at concentration C2, *L. minor* and *L. minuta* were characterised by an almost identical biomass increase (t = -0.0772, df = 4, p = 0.942).

In relative numbers however, the relative growth rate of *L. minor* did not differ significantly compared with *L. minuta* (all *p*-values > 0.01), even at concentration C4 ($t = 2.7358$, $df = 4$, $p = 0.052$). In short, the RGR did not result in a significant difference at a single concentration level and is, therefore, not able to differentiate between *L. minor* and *L. minuta*.

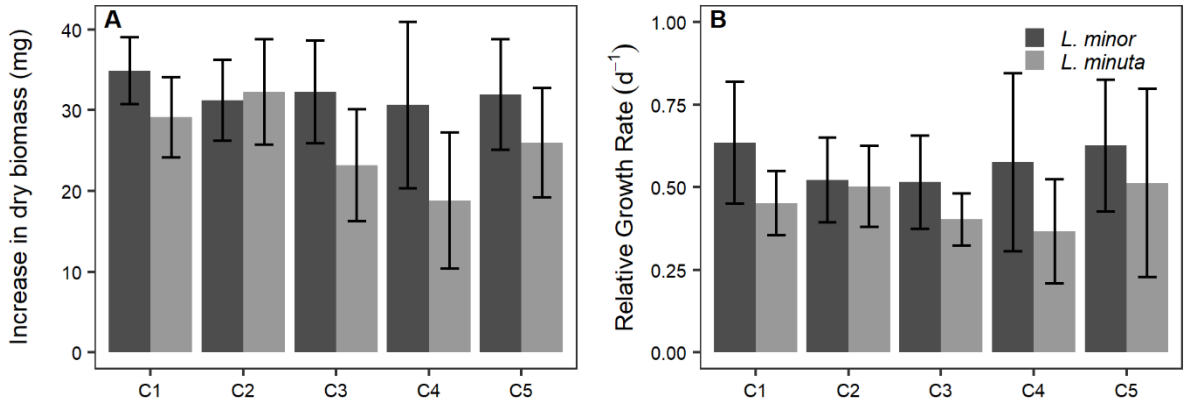

Figure 8.5: Change in biomass for* L. minor *(dark grey bars) and* L. minuta *(light grey bars). *A: absolute increase in biomass dry weight (mg) starting from day 2 (estimation) until day 4. B: Relative Growth Rate (RGR, d^{-1}) in a period of two days. Concentrations range from high (C1) to low (C5).*

8.3.3 Nutrient decrease versus biomass increase

Throughout the four day experiment, *L. minor* removed a maximum total amount of 2.1 mg nitrogen, while *L. minuta* removed 1.7 mg nitrogen (see also Figure 8.2), resulting in an approximated maximal average nitrogen removal rate of 0.525 $mg \cdot d^{-1}$ and 0.425 $mg \cdot d^{-1}$, respectively. Therefore, biomass-based nitrogen uptake rates were situated in between 2.1 $mmol \cdot g^{-1} \cdot d^{-1}$ (lowest observed dry weight of 17.6 mg) and 0.8 $mmol \cdot g^{-1} \cdot d^{-1}$ (highest observed dry weight of 49.1 mg) for *L. minor* and in between 1.5 $mmol \cdot g^{-1} \cdot d^{-1}$ (lowest observed dry weight of 20.2 mg) and 0.6 $mmol \cdot g^{-1} \cdot d^{-1}$ (highest observed dry weight of 47.7 mg) for *L. minuta*. Similarly, phosphorus was removed at a maximal average removal rate of 0.19 $mg \cdot d^{-1}$ and 0.25 $mg \cdot d^{-1}$ for *L. minor* and *L. minuta*, respectively. Resulting biomass-based phosphorus removal rates were situated between 0.4 $mmol \cdot g^{-1} \cdot d^{-1}$ and 0.1 $mmol \cdot g^{-1} \cdot d^{-1}$ for both *Lemna* species.

Nutrient removal in function of biomass increase (i.e. BBNR) varied between 20 $mg \cdot g^{-1}$ and 65 $mg \cdot g^{-1}$ for nitrogen and between 6 $mg \cdot g^{-1}$ and 30 $mg \cdot g^{-1}$ for phosphorus and combined the fluctuations in nutrient removal and biomass increase. In seemingly all cases a higher nutrient removal per gram newly formed biomass was observed for *L. minor*, though no significant differences were observed (all *p*-values > 0.01) (see Figure 8.6 and Table 8.3).

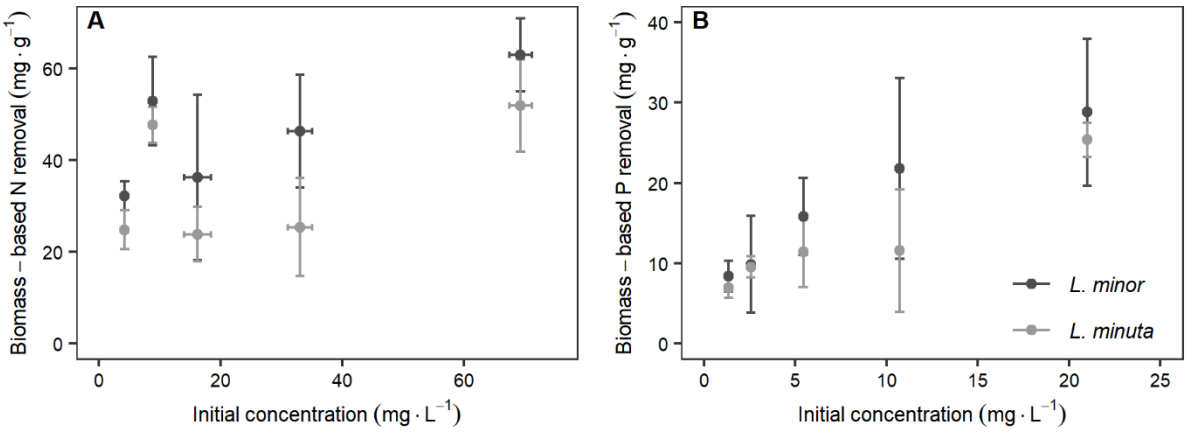

Figure 8.6: Nitrogen (A) and phosphorus (B) removal per gram newly formed biomass (dry weight) after four days for* L. minor *(dark grey circles) and* L. minuta *(light grey circles). *Similar patterns as in Figure 8.2 and Figure 8.3 can be observed, though differences between both* Lemna *spp. are influenced by the increase in biomass (see Figure 8.5).*

In short, BBNR observed similar differences in nutrient removal between *L. minor* and *L. minuta* as the FR, though it was not as powerful considering that all *p*-values were higher than the statistical threshold value (α = 0.01). A summary of the nutrient concentrations and obtained *p*-values for each of the considered functional traits is provided in Table 8.3.

Table 8.3: Nutrient concentrations and obtained* p*-values for three functional traits. *Results show minor similarities among the three functional traits measured for* L. minor *and* L. minuta. *Significant differences in functional traits (*p < *0.01) are underlined and were only observed at the nutrient level (i.e. FR). FR: Functional response; RGR: Relative growth rate and BBNR: Biomass-based nutrient removal.*

		C1	C2	C3	C4	C5
Concentration						
Nitrogen (mg·L⁻¹)		69 ± 2	33 ± 2	16 ± 2	8.8 ± 0.5	4.2 ± 0.1
Phosphorus (mg·L⁻¹)		20.99 ± 0.09	10.7 ± 0.1	5.43 ± 0.07	2.58 ± 0.03	1.33 ± 0.01
Functional traits (*p*-values)						
FR	Nitrogen	0.520	0.156	1.000	<u>0.003</u>	0.034
	Phosphorus	0.563	0.520	<u>0.002</u>	0.438	0.056
RGR		0.110	0.790	0.220	0.052	0.620
BBNR	Nitrogen	0.088	0.062	1.000	0.046	0.260
	Phosphorus	0.190	0.280	0.016	0.026	0.540

8.4 Discussion

8.4.1 Nutrient removal

Overall net nutrient removal by *Lemna minor* was higher than the nutrient removal exerted by *Lemna minuta* and contradicted the expectations of the latter having a higher functional response than the native *L. minor*. Even after removal of potential extreme values (three in total), no additional significant differences were observed. Furthermore, the difference in nutrient removal was also noticed when considering relative nutrient removal, showing that at low nutrient concentrations both species were efficient in using the provided nutrients. This efficiency decreased with increasing concentrations, though in general, *L. minor* illustrated a higher resource use efficiency. These results were not in line with field observations of *L. minuta* dominating *L. minor* in Belgian water bodies.

A similar contrast between field observations and experimental results was obtained when comparing two subspecies of the macrophyte *Phragmites australis*. Mozdzer *et al.* (2010) clearly observed the expected pattern of higher nutrient removal by the alien subspecies, but, when applied in practice, Rodríguez and Brisson (2015) observed a slightly higher nutrient removal by the native subspecies, especially towards phosphorus removal efficiency. According to Rodríguez and Brisson (2015), this discrepancy was related to the higher root biomass of the native *P. australis*, allowing it to take up more nutrients. This confirms both the obtained observations and reported findings of *L. minor* having longer roots (Njambuya *et al.*, 2011), and supports the vital role of roots in nitrogen uptake by *L. minor* as highlighted by Cedergreen and Madsen (2002). Additionally, these contrasting findings underline the idea that a clear difference between phylogenetically related species is hard to find and that further development and knowledge of appropriate testing methods is recommended. For instance, Colin *et al.* (2016) already mentioned the potential in applying biomarkers for identifying differences between native and invasive alien species at the sub-individual level, but also recognised the currently existing knowledge gap inhibiting its widespread application.

These results suggest that, despite its shown applicability at higher trophic levels (i.e. predator-prey interactions, see Dick *et al.* (2013)), the functional response approach does not show a higher nutrient removal by the known alien invader and therefore, does not allow to predict the invasive potential of *L. minuta*, solely based on nutrient removal. In combination with the contrasting results when comparing *Phragmites australis* (Mozdzer *et al.*, 2010; Rodríguez and Brisson, 2015), the functional response approach does not seem to be an appropriate method in predicting the invasiveness of alien macrophytes.

8.4.2 Biomass increase

In general, no significant differences were found in both absolute and relative biomass production between native and invasive *Lemna* plants. Similar to the functional response, extreme value removal (eight in total) did not result in additional significant differences with respect to the RGR. Still, *L. minor* performed better than *L. minuta,* except for condition C2, where an almost similar biomass increase was observed. This is in line with the higher observed nutrient removal by *L. minor* described in previous section, suggesting an overall higher efficiency in nutrient uptake by *L. minor*.

Relative growth rates (RGR) during the experimental period ranged from 0.5 d^{-1} to 0.6 d^{-1} for *L. minor* and from 0.4 d^{-1} to 0.5 d^{-1} for *L. minuta*. These values are higher than reported RGRs of duckweed, which are situated around 0.1 d^{-1} (Körner and Vermaat, 1998; Njambuya *et al.*, 2011) up to 0.3 d^{-1} (Cedergreen and Madsen, 2002; Gérard and Triest, 2014). This might be related to their applied test duration of 14 to 20 days, potentially leading to overcrowding and related decrease in growth rate (Driever *et al.*, 2005). In contrast, Körner and Vermaat (1998) only applied a duration of 3 days and observed a similarly low RGR of 0.1 d^{-1}. Yet, they used domestic wastewater as a growth medium, which differs from an ideal growth medium as defined by the OECD guidelines.

The observed RGRs suggest that *L. minor* is more effective in creating new (dry) biomass. However, when focusing on fresh weight (see Appendix, Table E.5 and Table E.6), the overall fresh biomass increase is larger for *L. minuta* than for *L. minor* (639 mg versus 406 mg of fresh weight, respectively), but a lower dry weight content reduces this difference (34 mg versus 31 mg of dry weight, respectively). Despite the lack of clear significant differences in RGR on a dry weight basis, *L. minor* might still be suppressed by *L. minuta* producing more new, fresh biomass with a lower dry weight content. This difference indicates an important drawback of using RGR for dominance prediction because some field-related information is not taken into consideration. For instance, Henry-Silva *et al.* (2008) investigated three different aquatic weeds and observed that RGR on a dry weight basis did not suffice to accurately predict infestation potential, suggesting to complement the RGR data with biomass density.

In general, no competitive superiority could be derived from the performed experiments. Moreover, the obtained results underline the fact that comparing RGRs of monocultures only depicts the potential direct competition and neglects more important indirect competition and interactions on the long run (Trinder *et al.*, 2013). Consequently, the relative growth rate provides information on biomass-based competition and dominance (Henry-Silva *et al.*, 2008), though is insufficient to describe or predict the invasive potential of macrophytes as no significant differences in RGR were observed.

8.4.3 Nutrient decrease versus biomass increase

Biomass-based nitrogen removal rates of both *Lemna* spp. fluctuated between 0.6 mmol·g^{-1}·d^{-1} and 2.3 mmol·g^{-1}·d^{-1} and, thereby, included the range observed by Cedergreen and Madsen (2002) for *L. minor* (0.6 mmol·g^{-1}·d^{-1} up to 0.9 mmol·g^{-1}·d^{-1}). Higher maximal nitrogen removal rates were obtained by *L. minor* when compared to *L. minuta*, which might be related to the observation of *L. minor* plants having longer roots, potentially increasing their nutrient uptake (Cedergreen and Madsen, 2002). Additionally, this difference in nutrient uptake was amplified by a higher net increase in biomass of *L. minuta* when compared with *L. minor* (see Appendix, Table E.3 and Table E.4), resulting in a difference in biomass-based nutrient removal rate in favour of *L. minor*.

Even so, under the assumption that *Lemna* spp. reallocate nutrients for biomass increase rather than biomass enrichment (Körner and Vermaat, 1998), nitrogen contents of both *L. minor* and *L. minuta* (ranging from 20 to 63 mg·g^{-1}) were comparable to the values obtained by Körner and Vermaat (1998), being 18.5 mg·g^{-1} up to 56.5 mg·g^{-1}, but were higher than reported by Cedergreen and Madsen (2002), being 5.6 mg·g^{-1} up to 27.3 mg mg·g^{-1}. In contrast, phosphorus content of both *Lemna* spp. (ranging from 6 mg·g^{-1} up to 30 mg·g^{-1}) was observed to be higher than reported by Körner and Vermaat (1998), being 3.6 mg·g^{-1} up to 7.2 mg·g^{-1}, which might be related to a difference in phosphorus content of the growth medium (1 mg·L^{-1} up to 21 mg·L^{-1} versus 1 mg·L^{-1} up to 14 mg·L^{-1}, respectively). Duckweed is known to be a P-hyperaccumulator and to store phosphorus as a precaution to future depletion (Gérard and Triest, 2014), which explains the increase in phosphorus removal at higher initial concentrations (see Figure 8.6). Nevertheless, biomass-based nutrient removal remains higher for *L. minor*, suggesting that *L. minor* requires more nutrients to produce new fronds (i.e. higher nitrogen and phosphorus content), while *L. minuta* biomass consists of more water. This is also supported by the observation of higher dry weight content of *L. minor* when compared to *L. minuta*.

In short, BBNR provides information about the efficiency of nutrient uptake per unit biomass, but lacks the ability to discriminate native from invasive alien species. Observed differences between both species were only marginally significant at the individual concentration level and were non-significant when accounting for multiple testing. Therefore, similar to FR and RGR, BBNR is not recommended to be used as the only technique to determine invasive potential, despite combining nutrient uptake and biomass increase.

8.4.4 Individual traits versus ecosystem-based techniques

Combining nutrient removal (input) and biomass increase (output) did not allow to clearly differentiate between the native *L. minor* and alien *L. minuta*. Overall, when looking at all three approaches, only two conditions were considered to be significantly different (see Table 8.3). Only the functional response showed a significant difference in phosphorus at concentration C3 and nitrogen at concentration C4. Firstly, this suggests that the FR is more sensitive towards differences between species, while the RGR is the least sensitive. In other words, differences are easier to be observed at the input-level than at the output-level.

Secondly, the differences between *L. minor* and *L. minuta* are clearer at lower nutrient concentrations, and require further research, while the absence of significant differences at high concentrations (C1 and C2) suggests that *L. minor* and *L. minuta* have a similar nutrient removal and biomass increase. Based on these individual specific traits, the invasive character of *L. minuta* could not be confirmed as *L. minor* displayed a higher nutrient removal and a higher relative growth rate. Consequently, taking into account *L. minuta*'s alien origin, the increasing in-field observations and its classification as 'widespread with moderate impact', the applied methods were considered to be insufficient for predicting a macrophyte's invasive potential. Nevertheless, the combined information provided by the individual traits (nutrient use and wet biomass increase) insinuated the presence of dominant behaviour of *L. minuta*, though this was not confirmed by the BBNR approach due to a highly fluctuating biomass increase.

Invasiveness is rarely determined by a single functional trait, but rather by a combination of traits and factors (Thuiller *et al.*, 2006; van Kleunen *et al.*, 2010). These factors include, among others, meteorological conditions, climate, resource availability of current environment, community complexity, frequency of disturbances, phenotypic plasticity, evolutionary adaptation and predator size (see for instance, Alpert *et al.* (2000), Baldy *et al.* (2015)), Gioria and Osborne (2014), Levine *et al.* (2003) and Riis *et al.* (2012). Therefore, experiments applying the FR, RGR or BBNR to determine a macrophyte species' invasive behaviour, should be supplemented with more complete and more complex ecosystem-scale research (e.g. Kovalenko *et al.* (2010)). Additional attention can be given to look for appropriate biomarkers not only to study the differences between closely related species at sub-individual level, but also to increase knowledge about the existing pathways and reactions to stress. As such, both policy makers and managers can be supported by data reflecting natural conditions more accurately instead of relying on the FR, RGR or BBNR to investigate the performance of different macrophyte species with respect to nutrient removal and biomass increase.

8.4.5 Contribution to the study objective

The aim of this chapter was to determine species-specific results for the functional response and relative growth rate and to evaluate their applicability towards predicting the invasive behaviour of an alien species. Forecasting the invasive behaviour of an alien species is crucial to develop proactive management plans by scoring or classifying alien species conditional to the discrepancy in functional traits. Moreover, the approach can be extended to the classification of native species and allows for an overall ranking of all species that are expected to occur. By avoiding the introduction of invasive species (both native and alien), higher species richness can be obtained in the managed system. Therefore, the applicability of this approach within the study objective (see Section 1.2.1) was tested with two *Lemna* spp., as these prefer eutrophic conditions and are known to occur as floating mats in ditches, ponds and wetlands (Janse and Van Puijenbroek, 1998).

Nutrient uptake and relative growth rate did not show to differ between *Lemna minor* and *Lemna minuta* (see Figure 8.2 and Figure 8.3) and suggested that both species display a similar invasive behaviour. More specifically, it can be hypothesised that both *Lemna* spp. provide a similar functionality after being introduced and affect the prevailing processes in a comparable way. However, further testing of additional traits and at ecologically relevant nutrient concentrations is required to confirm these observations. Nevertheless, the experimental results indicated that the suggested traits are insufficient to infer invasive behaviour, as the alien *L. minuta* has been observed to suppress the native *L. minor* under field conditions (Ceschin *et al.*, 2016). The selected traits might still detect significant discrepancies under different conditions, although they are not considered to be universally applicable.

The inability of the selected traits to confirm field observations (i.e. the alien *L. minuta* suppressing the presence of the native *L. minor*) reduced their overall value towards the development of proactive management plans. Yet, they still provide useful information on species-specific characteristics and are relatively easy to implement and follow up. Still, additional alternative traits can be considered to complement the selected resource-use efficiency and relative growth rate (e.g. growth form, specific leaf area, root length (Pérez-Harguindeguy *et al.*, 2013)), while including specific attention towards trait plasticity. The latter represents the ability to respond to stressors, which is often considered to be high in invasive species (Davidson *et al.*, 2011; Fagúndez and Lema, 2019). Moreover, it is often a driving factor in determining species richness within communities, as illustrated by Berg and Ellers (2010) and Barbour *et al.* (2019). By extending species-specific trait matrices with absolute trait values and trait-specific plasticity scores, a multivariate basis for species classification is created. Based on this classification, strategies can be developed to avoid the introduction of the most-invasive species in order to support a biodiverse ecosystem.

8.5 Conclusion

One input-based and one output-based approach were applied and supplemented with a third combined approach to test their applicability for predicting the invasive behaviour of the alien *Lemna minuta* when compared to the native *Lemna minor*. The FR approach did not meet the expectations of a higher resource removal by the invasive alien species, as it was observed that *L. minor* removes more nutrients than *L. minuta*, with significant differences at low nutrient concentrations. The net dry biomass increase was higher for *L. minor* at low nutrient concentrations, though no significant differences were observed when comparing the RGR of both species. In contrast, the increase in fresh weight was higher for *L. minuta*, which supported field observations of *L. minuta* dominating *L. minor*. As such, despite not meeting the expectations of a higher FR and RGR, the low nutrient requirement and high fresh weight increase supported the idea of *L. minuta* being more invasive than *L. minor*. In the observed range, no dominance of the invasive alien macrophyte could be clearly inferred by applying a single approach, suggesting that other functional traits (e.g. temperature resistance, germination period, ...) or environmental conditions (e.g. seasonality, solar radiation) might provide a competitive advantage (Riis *et al.*, 2012). Therefore, it is recommended to supplement currently existing functional traits with more in-depth and ecosystem-based research as the former, when applied individually, lacks the ability to identify and predict an invasive alien species with a moderate impact.

9

Effects of partial harvesting and species invasion on biomass production[7]

Highlights

- Biomass production of host species is not affected by invasion
- Growth rate is positively affected by biomass removal
- Overcrowding negatively affects growth rate
- Invasive *L. minuta* shows to dominate over native *L. minor*

[7] This chapter is based on Van Echelpoel, W.; De Troyer, N. and Goethals, P. L. M. (in preparation) Effects of species invasion and repetitive partial removal on the interaction between *Lemna* spp

Abstract

Increasing globalisation and ongoing climate change threaten biodiversity with rising rates of biological invasions globally, being introduced either accidentally or intentionally. Invasion prevention and impact containment are therefore imperative when developing freshwater management schemes. In this study, monocultures of two duckweed species (the native *Lemna minor* and the alien *Lemna minuta*) were exposed to nine different scenarios combining removal frequency ('none', 'low' and 'high') and biomass introduction frequency ('none', 'low' and 'high'). Biomass removal was considered to be non-specific, while only biomass of the opposing species was introduced to not directly affect the original host species. Experiments were run for 34 days, consisting of four days acclimation, twenty-four days of management and six days of undisturbed growth. The results illustrate that the overall growth rate was slightly higher for *L. minuta* compared to *L. minor* (0.116 d^{-1} versus 0.111 d^{-1}) and time-dependent, showing to decrease in time due to overcrowding. During the treatment period, biomass of the host species increased and showed a diverging behaviour among scenarios. Afterwards, discrepancies in biomass dry weight decreased, while the production of primary species showed to be unaffected by the introduction of a second species. Consequently, with total biomass benefitting from species introduction, dominance by the host species decreased in time and plateaued towards the end of the treatment period. Nevertheless, higher growth rates for *L. minuta* supported higher biomass ratios with *L. minuta* as host species compared to biomass ratios with *L. minor* as host species. This indicates that assessment of the introduction frequency prior to biomass removal is crucial to avoid the detrimental effects of invasive species, making the decision on management actions and frequency highly case-dependent. Hence, additional studies are essential to extend the presented findings towards a comprehensive characterisation of the interaction between management and natural processes.

9.1 Setting the scene

In Chapter 8, the applicability of trait-based assessment was studied to forecast the invasive behaviour of an alien species prior to its introduction. Despite having value towards protecting the local native biodiversity and understanding the species' dispersal dynamics, this suite of pre-introduction studies provides limited additional value when an alien species is already established. To avoid further dispersal, colonisation and biodiversity loss, it is imperative that invasive alien species (IAS) and their interactions with land conversion, hydrological alterations and climate change are identified and contained (Alexander *et al.*, 2014; Richter *et al.*, 2003).

Conceptually, a series of steps occurs during biological invasion, the identification of which helps illustrating the invasion process, improving the interpretation of results and inventing the appropriate management plans (Colautti and MacIsaac, 2004). In short, introduction of non-indigenous species (NIS) requires the transport of propagules by an abiotic (e.g. wind, runoff) or a biotic (e.g. pollination, international shipping) vector (Murphy *et al.*, 2019), which represents a first barrier in the invasion process. Subsequently, in the presence of suitable abiotic conditions and relatively low biotic resistance (i.e. a second and third barrier), the NIS has the potential to establish successfully. Lastly, both disturbance frequency and intensity determine the survival of the introduced species and whether the invaded area will act as a sink area or a new hotspot for high-density colonisation and local dispersal. Despite the existence of these barriers, no 'one-method-fits-all' exists to efficiently tackle biological invasions. For instance, border control increases the first barrier, though is logistically challenged due to the high degree of globalisation. Similarly, biodiverse communities have the capacity to slow down invasion and mitigate negative impacts, but the effectiveness of this biotic resistance depends highly on the exerted propagule pressure, prevailing resource dynamics and degree of niche occupancy (Davis *et al.*, 2000; Levine *et al.*, 2004). Therefore, increased attention towards successful eradication measures for established invasive alien species is vital and further supported by the continuously growing list of acknowledged harmful IAS (IUCN, 2019).

Unfortunately, complete eradication of IAS is hard, costly and often harmful towards non-target species (Myers *et al.*, 2000). Literature on success stories is sparse but increasing, though represents a bias towards insular systems and terrestrial animals (Simberloff *et al.*, 2018; Zavaleta *et al.*, 2001). Moreover, the majority of eradication programs remains unpublished or hidden in grey literature due to observed failures caused by incomplete elimination, continued introduction by a nearby species pool and range shifts due to climate change (Rahel and Olden, 2008). Hence, in order to embrace and counter these eradication challenges, an integrated spatiotemporal dynamic approach and follow-up is required.

Moreover, aside from being challenged by a variety of abiotic and biotic factors, a disturbance is generated when eradicating an IAS, creating an opportunity for competitive species to establish or colonise. This is considered beneficial when extracting an invasive alien species from an area under native propagule pressure, though has a potential detrimental effect when a native species is affected. Both pressures (extraction frequency and incoming propagules) are expected to interact and affect biomass production, creating new and unfamiliar ecosystems that require dynamic and feedback-oriented management plans. For instance, adopting repetitive management with partial harvesting reduces the required funding per intervention, decreases the dispersal potential of IAS and allows to continuously update management based on intermediate results (Myers *et al.*, 2000). Moreover, it creates less disturbance and provides an opportunity for native species to establish and compete with the IAS (Catford *et al.*, 2009), following natural or artificial introduction. Unfortunately, the current lack of appropriate guidelines on the frequency and intensity of these partial interventions impedes their (successful) implementation.

These challenges illustrate the need for alternative eradication activities, especially because the impacts of alien species on ecosystem structure and functioning remain highly species-specific. For instance, the alien *Lemna minuta* is invasive in Belgium, causing a moderate impact on the abiotic and biotic conditions within surface waters. More specifically, similar impacts are observed for *L. minuta* and the native *L. minor*, as their presence in aquatic systems is often characterised by dense mats that negatively affect aquatic life underneath by decreasing light penetration and oxygen concentration (Janes *et al.*, 1996; Janse and Van Puijenbroek, 1998). Consequently, removal of these mats is beneficial to (1) improve light penetration and (2) reduce local stock of the invasive *L. minuta* (Ceschin *et al.*, 2016). However, as most *Lemna* spp. reproduce in a vegetative way (Hillman, 1961), complete eradication programs without follow-up tend to be ineffective as a single frond is sufficient to restart colonisation.

Within this chapter, attention is given to the temporal trend of primary production under a combination of two external pressures: partial biomass removal and biomass introduction. The aim is to determine if biomass production is affected and whether a native population responds differently compared to an alien population. To do so, both *L. minor* and *L. minuta* are exposed to (i) three levels of biomass removal: none, low frequency and high frequency and (ii) three levels of propagule pressure by the opposite species: none, low frequency and high frequency. By tackling these issues, an answer is provided to RQ3.2, as defined in Chapter 1. Hence, this chapter concludes with a statement on how monocultures are affected by artificial removal in combination with natural introduction of a competitor.

9.2 Materials and Methods

9.2.1 Experimental setup

Duckweed (*Lemna minor* and *Lemna minuta*) were collected in a ditch in Ghent (51.055111, 3.685639) and separated in the lab to grow monocultures of *L. minor* and *L. minuta*. Stock cultures were grown in plastic aquaria containing 3 L of nutrient medium based on OECD and ISO guidelines for chemical testing with *L. minor*, which is referred to as the full strength modified Steinberg medium (OECD, 2006) described in Table 9.1. Fluorescence lamps were used to provide 16 hours of light, followed by 8 hours of darkness, with an intensity at water surface of 36 $\mu mol \cdot m^{-2} \cdot s^{-1}$ up to 55 $\mu mol \cdot m^{-2} \cdot s^{-1}$ (average: 44 ± 5 $\mu mol \cdot m^{-2} \cdot s^{-1}$). Water temperature was registered continuously and varied between 16.9 °C and 20.5 °C (average: 18.6 ± 0.5 °C). Every six days new medium was provided and aquaria were rinsed thoroughly with tap water.

The experiment entailed a full-factorial design including three levels of introduction (none, low, high) and three levels of removal (none, low, high), resulting in a total of nine scenarios (Figure 9.1). Each scenario was repeated three times and applied to each *Lemna* species, providing a total of 54 containers. Tests started with a single species, henceforth referred to as 'primary species', and were complemented (if applicable) with the competing species, referred to as 'secondary species'.

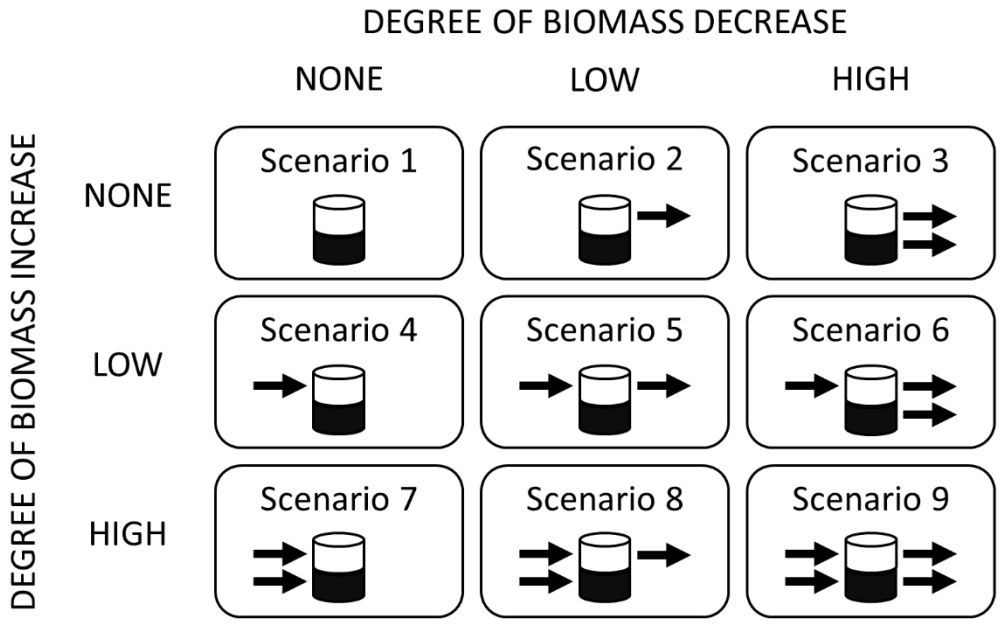

Figure 9.1: Experimental set-up for the assessment of invasion vulnerability. *The different scenarios account for three levels of biomass increase (e.g. due to dispersal of competitor) and biomass decrease (e.g. due to management, herbivory or dispersal). Arrows indicate the intensity of introduction (towards aquaria) and removal (from aquaria), while cylinders indicate the aquaria filled with 3 L (black colour) of medium. Each scenario is repeated three times and applied to two different species.*

Each container held 3 L of diluted OECD medium with increased phosphorus content as described in Table 9.1. Floating separators were introduced to all containers for pragmatic reasons. More specifically, they allow nutrient concentrations to be the same for both species and to avoid the mixing of the two species, which eases temporal follow-up of the produced biomass. However, the created separation does not allow for physical interaction between the individuals of different species, which causes a loss of relevance towards natural conditions. Light and temperature conditions remained unaltered. Throughout the experiment, containers were randomised every 2 days and water loss due to evapotranspiration was compensated by adding deionised water to maintain a volume of 3 L. To avoid excessive algae growth and nutrient depletion, the nutrient medium was replaced every six days.

Table 9.1: Composition of test medium used for growing monoculture (Full strength) and performing the test (Reduced). *Composition is based on the Steinberg medium used for chemical testing with* Lemna minor *(OECD, 2006).*

	Full strength	**Reduced**
Macronutrients	**(mg·L^{-1})**	**(mg·L^{-1})**
KNO_3	350	70.0
KH_2PO_4	30	9.0
K_2HPO_4	4.2	1.26
$MgSO_4$	49	9.8
$Ca(NO_3)_2$	205	41.0
Micronutrients	**(µg·L^{-1})**	**(µg·L^{-1})**
H_3BO_3	120	24
$ZnSO_4$	100	20
Na_2MoO_4	40	7.7
$MnCl_2$	130	26
$FeCl_3$	456	91
Na-EDTA	1500	300

At the start of the experiment, 500 mg fresh weight of the primary species was introduced in the containers, followed by four days of undisturbed growth. Determination of the fresh weight was performed by collecting biomass on a sieve and blotting the fronds with tissue paper to extract attached water as much as possible. Introduction and removal actions were defined to occur every 4 (8) days in case of high (low) frequency, scheduled intermittently (Figure 9.2). A full cycle consisted of 8 days during which 2 (1) introduction and 2 (1) removal events occurred for the high (low) frequency containers. In total, three cycles were run, followed by six days of undisturbed growth. Introduction rates of the secondary species were fixed at 50 mg fresh weight (i.e. 10 % of initial biomass), while removal rates were set at 20 % of the total biomass.

Although both rates are relatively arbitrary, similar removal rates have been used in literature, see for instance Tang *et al.* (2017). Removal was designed to be non-specific, e.g. with 1.0 g of total biomass consisting of 80 % *L. minor* and 20 % *L. minuta*, a total of 0.2 g would be removed, combining 0.16 g of *L. minor* and 0.04 g of *L. minuta*. Due to the scheduling of these events, information on biomass wet weights was collected at a 2-day frequency. Moreover, every 4 days the dry weight of the removed biomass was determined after being dried at 60 °C for at least 48 h (OECD, 2006).

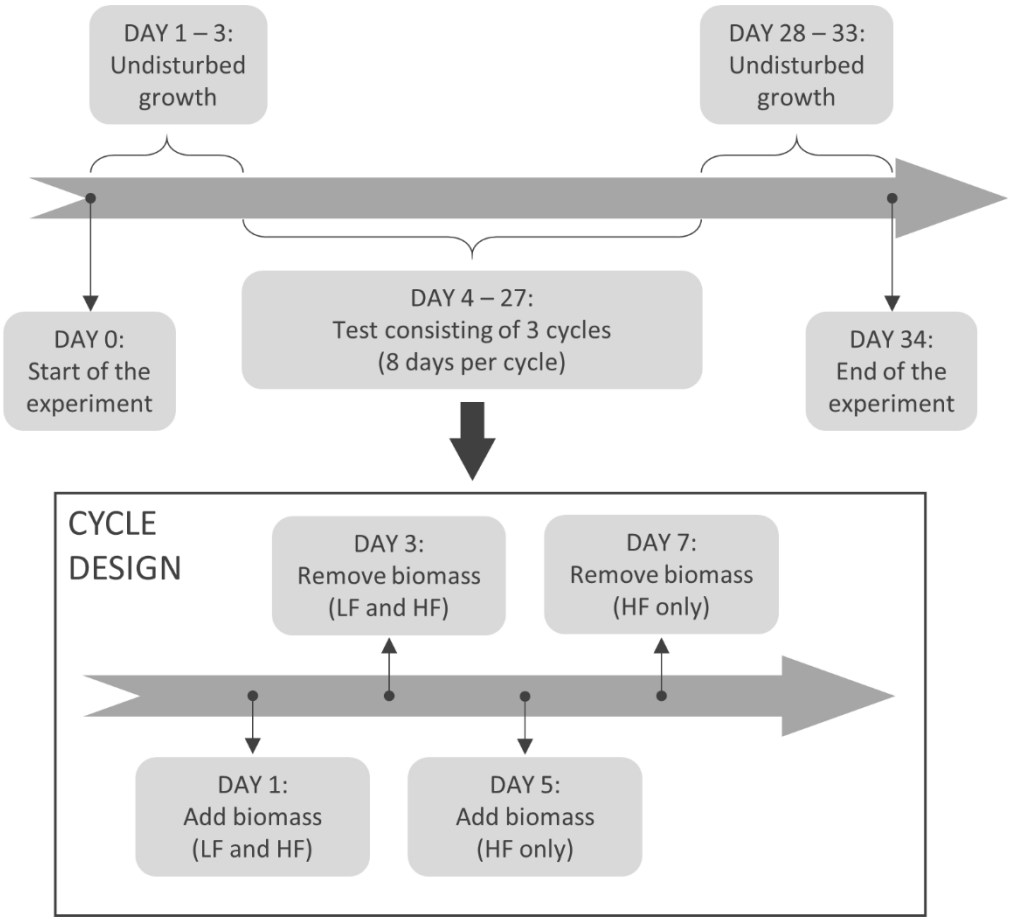

Figure 9.2: Schedule for implementation of different management and introduction scenarios in time. *Different scenarios have been defined (see Figure 9.1) and will experience different pressures. HF: high frequency; LF: low frequency.*

Simulations of *Lemna* spp. biomass over time were performed to assess the discrepancy between theory and practice. Components affecting biomass were (i) growth, (ii) introduction and (iii) removal. For each time point, the new biomass was calculated based on previous time point and the applicable management, with the calculation following Equation 9.1. Simulations for a range of relative growth rates are graphically depicted in Figure F.1.

$$M_i = e^{r \cdot t} \cdot M_{i-1} + m_i \cdot 0.05 - k_i \cdot 0.2 \cdot e^{r \cdot t} \cdot M_{i-1} \qquad \text{(Equation 9.1)}$$

With M_i representing the biomass at point i, r the relative growth rate of the considered species (d^{-1}), t the time between two sampling events (d), m_i reflecting whether or not biomass should be added (dichotomous) (-) and k_i reflecting whether or not biomass should be removed (dichotomous) (-).

9.2.2 Data analysis

Records of fresh weight collected throughout the test were converted into dry weight values by applying a species-specific dry weight ratio based on the final biomass. It was assumed that temporal changes in the dry weight ratio are negligible, following the similarity in resource provision. For each *Lemna* species, a single dry weight ratio was calculated. Subsequently, the resulting dry weight scores were used to determine the temporal trends in biomass, relative growth rate and relative dominance.

Statistical analysis of the final biomass relied on the assumption that the obtained values originated from a normal distribution. Normality was tested for by applying the Shapiro-Wilk test with Benjamini-Hochberg correction for multiple testing. Subsequently, homoscedasticity was checked for by performing a Bartlett-test. Although no significant differences from normality or homoscedasticity were observed (all $p > 0.05$; results not shown), results should merely be considered as support for visual observations instead of absolute values due to the low number of replicates (i.e. $n = 3$). Analysis of Variance (ANOVA) was used to indicate whether a significant difference in final biomass among scenarios was present and, if significant (i.e. $p < 0.05$), followed by two-sample Student's *t*-tests with Benjamini-Hochberg correction for multiple testing.

Temporal trends in biomass were assessed by means of generalised linear mixed effect models (GLMMs), considering introduction, removal and time as fixed effects and the aquarium as random effect. Saturated models included all interactions among the fixed effects and included a random intercept and an autoregressive variance-covariance structure. Model simplification focused on optimising the random effect structure, followed by stepwise exclusion of (interacting) fixed effects (Zuur *et al.*, 2009). Elimination of (interacting) variables decreased the variance during parameter estimation, yet increased bias. The Akaike Information Criterion (AIC) was used to decide on the in- or exclusion of an (interaction) effect and represents model fit, while penalising for complexity. More information on the development of these linear mixed effects models can be found in Appendix, Section F.2.

9.3 Results

Determination of final fresh and dry weight of each primary and secondary species provided an average dry-to-wet weight ratio of 0.053 ± 0.003 $g \cdot g^{-1}$ for *L. minor* and 0.051 ± 0.005 $g \cdot g^{-1}$ for *L. minuta* ($N = 27$), being confirmed by the observed range in literature (i.e. between 0.05 $g \cdot g^{-1}$ and 0.15 $g \cdot g^{-1}$) (Appenroth *et al.*, 2017; Cedergreen and Madsen, 2002). Slight differences in dry weight ratios occurred throughout the test period (see Appendix, Figure F.3), yet ratios determined on the overall final biomass were considered to be more relevant to convert the intermediate fresh weights. The resulting dry weight values were subsequently used for analysis of the final biomass production and temporal trends in biomass production and growth rates, and will be used from here on unless mentioned otherwise.

9.3.1 Biomass production

Final biomass of the primary species clearly differed among the nine scenarios for both *L. minor* ($F = 36.27$, $p < 0.001$; Table 9.2) and *L. minuta* ($F = 53.81$, $p < 0.001$; Table 9.2), showing to be highest when no removal was performed (Figure 9.3). Lower *L. minor* biomass was obtained when biomass was actively removed (Figure 9.3), though significant differences were only observed for a few cases (see Appendix, Table F.1). More specifically, in comparison to the control treatment, significantly lower final biomass scores were obtained under (i) no introduction (1.38 ± 0.10 g versus 0.71 ± 0.03 g; p = 0.024) and (ii) high-frequency introduction (1.28 ± 0.04 g versus 0.64 ± 0.04 g; p = 0.001) of *L. minuta*. Similar scenario-specific differences were observed for *L. minuta* (see Appendix, Table F.3), illustrating the significant effects of high-frequency removal on biomass production. Here, in comparison to the control treatment, significantly lower final biomass scores were obtained under (i) no introduction (1.30 ± 0.07 g versus 0.59 ± 0.08 g; p = 0.007) and (ii) high-frequency introduction (1.32 ± 0.10 g versus 0.66 ± 0.03 g; p = 0.017) of *L. minor*. Similar treatment effects were observed for the total biomass (see Appendix, Figure F.2).

Table 9.2: ANOVA results of final dry weight, grouped per scenario. *Nine scenarios were considered when the focus species is the primary species, while only six scenarios were considered in case the focus species was the secondary species. Differences tend to be more significant among groups when more scenarios are considered.*

Primary species	Focus species	# Scenarios	*F*-Statistic	*p*-value
L. minor	*L. minor*	9	36.27	$1.48 \cdot 10^{-9}$
	L. minuta	6	22.29	$1.09 \cdot 10^{-5}$
L. minuta	*L. minor*	6	18.13	$3.2 \cdot 10^{-5}$
	L. minuta	9	53.81	$5.32 \cdot 10^{-11}$

Introduction of a secondary species had a limited effect on the biomass of the primary species, with patterns being highly similar between the control and high-frequency introduction scenarios (see Figure 9.3). At low-frequency introduction, however, *L. minor* seems negatively affected by the introduction of *L. minuta* when no biomass is removed, providing lower biomass (0.90 ± 0.16 g) compared to (i) the control (1.38 ± 0.10 g; $p = 0.042$) and (ii) the high-frequency introduction scenarios (1.28 ± 0.04 g; $p = 0.080$). In contrast, *L. minuta* appears to be positively influenced by the introduction of *L. minor* when biomass is removed at a low frequency, producing more biomass (1.18 ± 0.06 g) than (i) the control (1.07 ± 0.08 g; $p = 0.184$) or (ii) high-frequency introduction scenario (0.86 ± 0.05 g; $p = 0.012$).

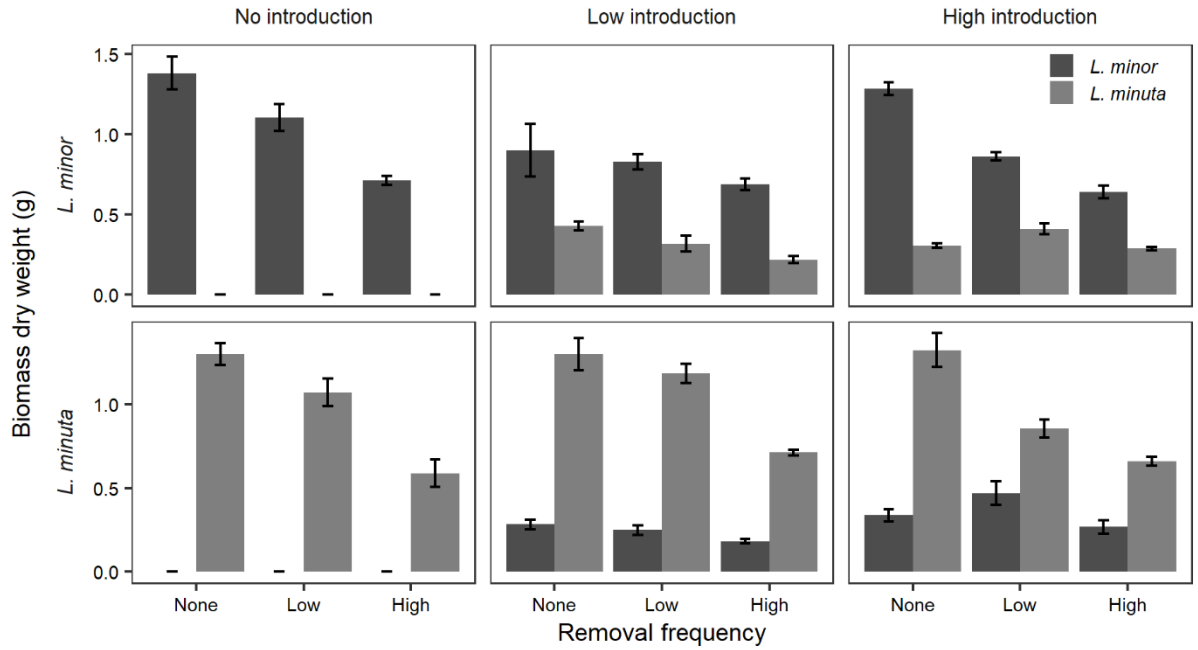

***Figure 9.3: Final biomass of* Lemna minor *and* Lemna minuta.** *Biomass dry weight of* L. minor *(dark grey) and* L. minuta *(light grey) was determined after 34 days exposure to 9 treatment scenarios. An effect of biomass removal is visible for each species, though is less clear for primary species in case of low introduction-frequency of a secondary species. Secondary species benefit from higher introduction rates, especially in combination with low-frequency removal.*

Secondary species experienced similar effects as primary species, with significant differences in biomass production among all considered scenarios ($p < 0.05$; Table 9.2). High-frequency removal caused final biomass of *L. minor* to be lower compared to the control treatment (i.e. no removal) under both low-frequency (0.18 ± 0.01 g versus 0.28 ± 0.03 g; $p = 0.059$) and high-frequency (0.27 ± 0.04 g versus 0.34 ± 0.04 g; $p = 0.119$) introduction.

Analogously, final biomass of *L. minuta* was lower under high-frequency removal compared to removal-free for both low-frequency (0.22 ± 0.02 g versus 0.43 ± 0.03 g; p = 0.006) and high-frequency (0.29 ± 0.01 g versus 0.31 ± 0.01 g; p = 0.136) introduction (Figure 9.3). Overall, no significant differences were obtained between low-frequency and high-frequency introduction (all p > 0.05), suggesting that within-species competition might have counteracted the elevated introduction rates. Remarkably, combining low-frequency removal and high-frequency introduction (i.e. scenario 8) caused the highest biomass production of both *L. minor* (0.47 ± 0.07 g) and *L. minuta* (0.41 ± 0.03 g), but only showed to be significantly higher for *L. minuta* compared to the control treatment (0.31 ± 0.01 g; p = 0.037) and high-frequency removal (0.29 ± 0.01 g; p = 0.037) scenario (Figure 9.3).

Relative growth rates based on the initial and final biomass were slightly higher for *L. minuta* (0.116 d^{-1} (s < 0.001 d^{-1})) compared to *L. minor* (0.111 ± 0.007 d^{-1}) in undisturbed environments, though did not show to be significantly different (t = -1.213, df = 2.008, p = 0.349) (see Table 9.3). Obtained rates were used for updating the applied growth rates within the simulations performed in Section 9.2.2 (i.e. Equation 9.1 and Figure F.1), though divergence between observations and simulations was expected, as obtained rates were relatively low compared to literature and reported in previous chapter (i.e. 0.1 d^{-1} up to 0.5 d^{-1} (Gérard and Triest, 2014; Njambuya *et al.*, 2011)), which insinuates the presence of time-specific growth rate fluctuations.

Table 9.3: Relative growth rates (RGRs) for* Lemna minor *and* Lemna minuta *based on the overall biomass increase during the test period (34 days). *Only scenarios supporting undisturbed growth (i.e. no biomass removal) of the primary species were considered for RGR calculation. Each scenario contained three replicates, which were used to determine an average, scenario-specific RGR and sd (standard deviation). An overall RGR was based on the average of the scenario-specific RGRs.*

Species	Scenario	RGR (d^{-1})	sd (d^{-1})
L. minor	1	0.116	0.002
	4	0.103	0.006
	7	0.114	0.001
	Mean	*0.111*	*0.007*
L. minuta	1	0.116	0.001
	4	0.116	0.002
	7	0.116	0.002
	Mean	*0.116*	*< 0.001*

9.3.2 Temporal patterns

9.3.2.1 Biomass

Biomass of the primary species increased in time and illustrated the effect of removal-based management on biomass production. Low-frequency (days 6, 14 and 22) and high-frequency removal (days 6, 10, 14, 18, 22 and 26) occasions occurred as minor drops in biomass (Figure 9.4), causing temporal biomass patterns to diverge. Towards the end of the experiment, biomass values tended to reach a plateau, with a seemingly higher effect for scenarios without biomass removal. For instance, without being exposed to biomass removal, *L. minor* biomass increases sharply at first (± 4 days), followed by a more gentle increase for a longer time period (i.e. 15 to 20 days), after which the increase in biomass remains low. Simultaneously, *L. minor* populations exposed to biomass removal follow a similar pattern, yet tend to keep growing during the third stage and thereby decrease the difference with the control treatment (Figure 9.4). A similar pattern can be observed for *L. minuta* as primary species, though shows a steeper increase during the first period while plateauing faster than *L. minor* (Figure 9.4).

The introduction of a secondary species did not seem to affect the observed patterns for *L. minor* and *L. minuta*. This solidifies the suggestion that invasion of a secondary species hardly affects the population dynamics of the primary species, as inferred from Figure 9.3. Moreover, Figure 9.4 illustrates the convergence of biomass patterns among different scenarios and indicates that more significant discrepancies occurred throughout the treatment period compared to Figure 9.3, while a higher similarity in overall biomass can be expected after a certain amount of time (i.e. hypothetical elongation of the applied time window).

The developed GLMMs confirmed that undisturbed growth occurred during the first time period, as no interactions with the performed treatment were included in the model. In contrast, during the treatment period, biomass production of the primary species was significantly affected by the applied treatment, including both individual and interactive effects (see Appendix, Table F.5 and Table F.6). More specifically, the inclusion of removal frequency showed to significantly improve model fit for both *L. minor* and *L. minuta* ($p < 0.001$) during the treatment period, while time-specific effects of introduction frequency were relatively non-significant ($p > 0.05$) (see Appendix, Table F.5 and Table F.6). Lastly, within the third period (i.e. undisturbed growth) a significant effect of treatment was observed for *L. minor*, while for *L. minuta* only the inclusion of removal frequency significantly improved model fit. GLMMs for *L. minor* showed to fit the observations relatively well, with a limited residual pattern in the temporal dimension (see Appendix, Figure F.6, Figure F.7 and Figure F.8). Similarly, GLMMs for *L. minuta* provided an acceptable fit, though showed a larger residual pattern within the temporal dimension (see Appendix, Figure F.9, Figure F.10 and Figure F.11).

Simulations (see Equation 9.1 and Figure F.1) greatly underestimated the obtained biomass throughout the test period (see Figure 9.4). Only final biomass predictions for undisturbed biomass growth (i.e. no removal) showed to be relatively accurate (especially for *L. minuta*, see Figure 9.4), mostly because the growth rate was based on these data points (see Table 9.3). In contrast, biomass predictions for low-frequency and high-frequency removal scenarios indicated an underestimation of the final biomass (Figure 9.4), due to applying a time-independent growth rate. Indeed, the discrepancy between observations and simulations gradually increased until around day 24 (*L. minuta*) or day 30 (*L. minor*), after which the difference became smaller (see Appendix, Figure F.4). Final biomass tended to be most accurately predicted when no biomass was removed, which contrasted with the highest errors observed during the previous time points (see Appendix, Figure F.4). This indicates that the applied biomass density throughout this test is already sufficient to influence the relative growth rate and that the time-independent RGR is an incorrect simplification to represent the growth dynamics of *Lemna* spp.

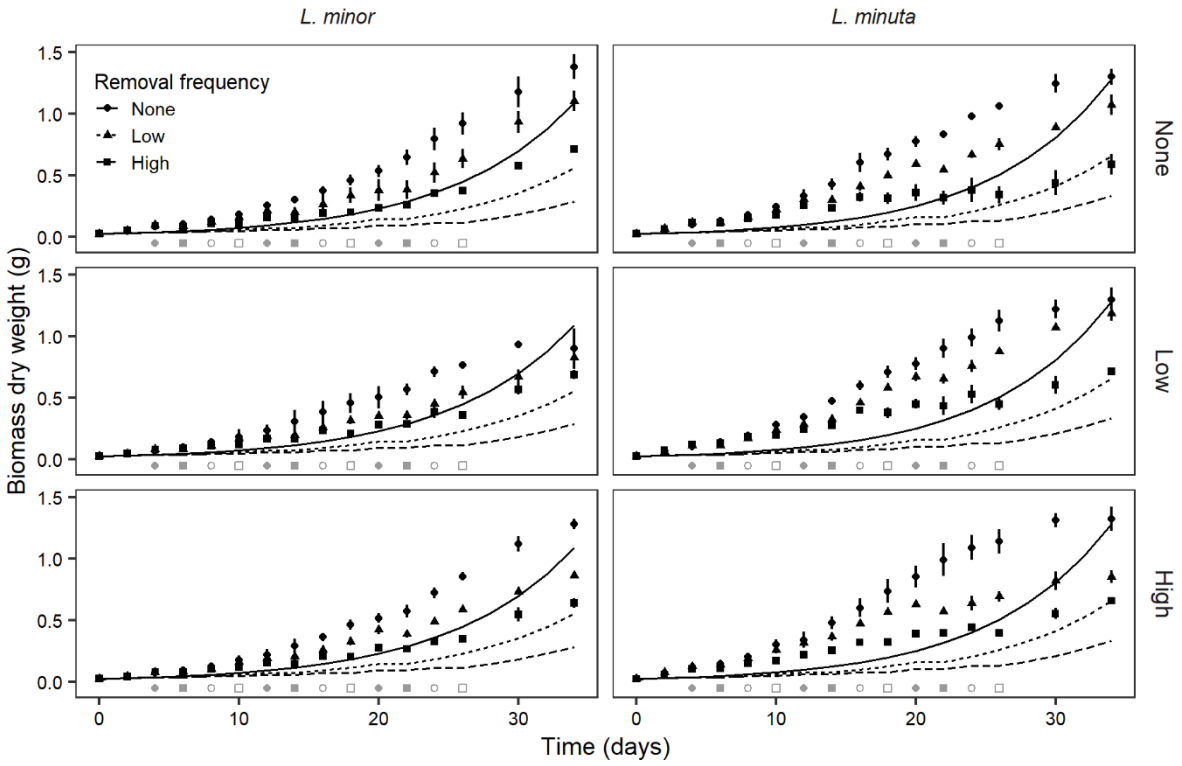

Figure 9.4: Temporal increase of biomass for the two primary species (columns) at three levels of introduction pressure (rows). *An increase in biomass is expected and observed in time, though a clear discrepancy exists between the simulations (lines) and observations (black symbols). Simulations rely on species-specific growth rates (*L. minor*: 0.111 d^{-1};* L. minuta*: 0.116 d^{-1}) and are unaffected by introduction of a secondary species, while no clear effect can be observed in practice. Grey symbols represent introduction (circles) and removal (squares) events, with filled symbols indicating the low frequency pressure.*

9.3.2.2 Relative growth rate and biomass ratio

Temporal assessment of the relative growth rate corroborated previous statements and indicated the dynamic character of the growth rate throughout the test period. In general, growth rates were highest directly after the start of the test and decreased in function of time (Figure 9.5). Initial growth rates for *L. minuta* were higher than for *L. minor*, though dropped faster to a similar rate from day 6 onwards. The highest drop in growth rate was observed for *L. minuta* at high-frequency introduction and low-frequency removal from 0.566 ± 0.045 d^{-1} (day 2) to 0.024 ± 0.010 d^{-1} (day 34), while smaller drops were obtained for *L. minor* (Figure 9.5). Subsequent growth rates showed to depend on removal frequency, with a slightly higher degree of stability when no biomass was removed, as illustrated by the drop in growth rate on day 8 due to biomass removal on day 6. Contrasting the effects of biomass removal, no effects of introduction pressure were observed (Figure 9.5).

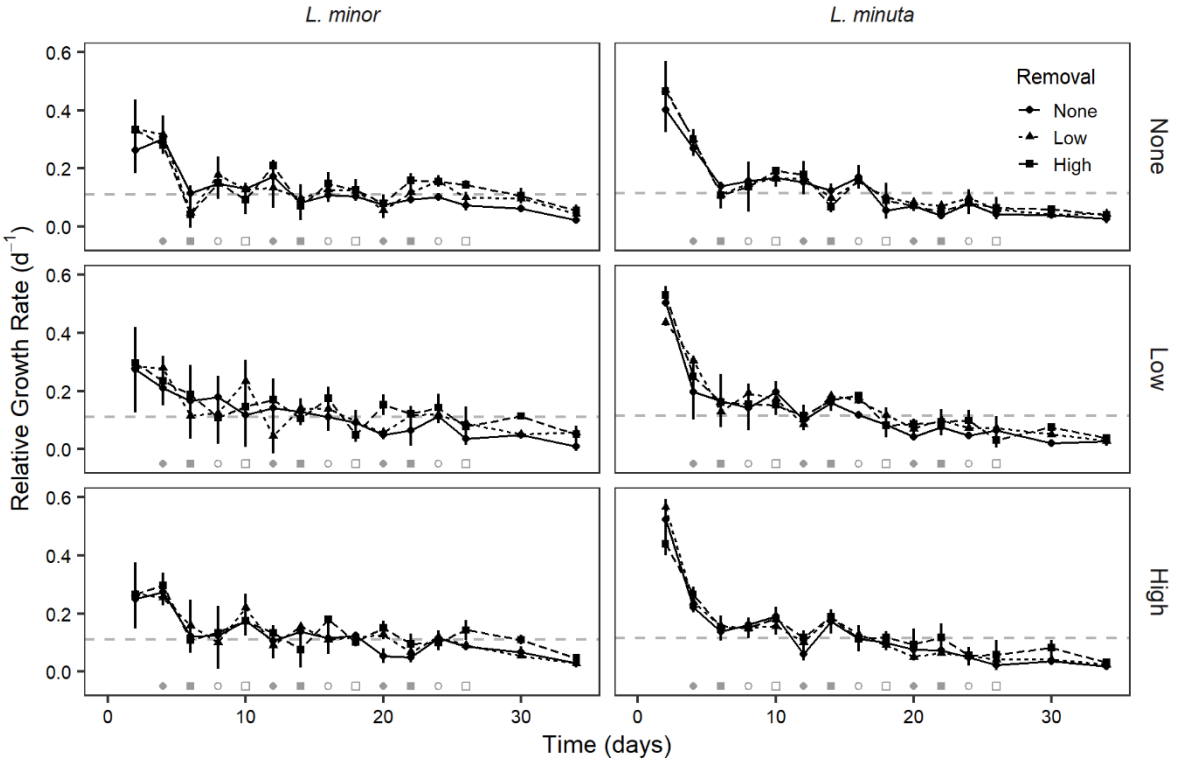

Figure 9.5: Temporal evolution of relative growth rate (RGR) for* Lemna minor *and* Lemna minuta *for different management scenarios. *A decrease in RGR is obtained for each primary species, indicating the temporal dynamics of the RGR and the incorrect assumption of using a stable RGR (dashed grey line) for simulations. Initial exceedance of the fixed growth rate causes higher reproduction at the beginning of the experiment, which propagates through time and results in underestimated biomass values (Figure 9.4). Grey symbols represent introduction (circles) and removal (squares) events, with filled symbols indicating the low frequency pressure.*

The relative dominance of the primary species over the secondary species exceeded the equilibrium condition (i.e. ratio = 1) throughout the whole test. Observations followed the expected decrease and indicated relatively high dominance during the first introduction event (day 4), followed by a decrease due to the exerted introduction and reproduction pressure. Ratios for the primary *L. minuta* were generally higher compared to *L. minor*, illustrating the effects of higher (overall and initial) growth rates (see Table 9.3 and Figure 9.5). Removal frequency limitedly affected the ratio, except for *L. minuta* under low-frequency introduction during the first weeks of the experiment (Figure 9.6).

Increased introduction frequency exerted a slightly negative effect on biomass ratios, illustrated by a faster drop for the primary *L. minor* at high-frequency introduction compared to low-frequency introduction (Figure 9.6). Similarly, relative dominance of *L. minuta* seemed to be slightly lower under high introduction pressure compared to being under low introduction pressure. Moreover, a minor effect of introduction frequency on final biomass ratio was observed for *L. minuta*, while the final ratio for *L. minor* remained relatively similar. This indicates that the introduction effect of the secondary *L. minuta* is less frequency-dependent than the introduction effect of *L. minor*, suggesting that the former is more assertive towards biomass production and confirming the (overall and initial) higher growth rates for *L. minuta*.

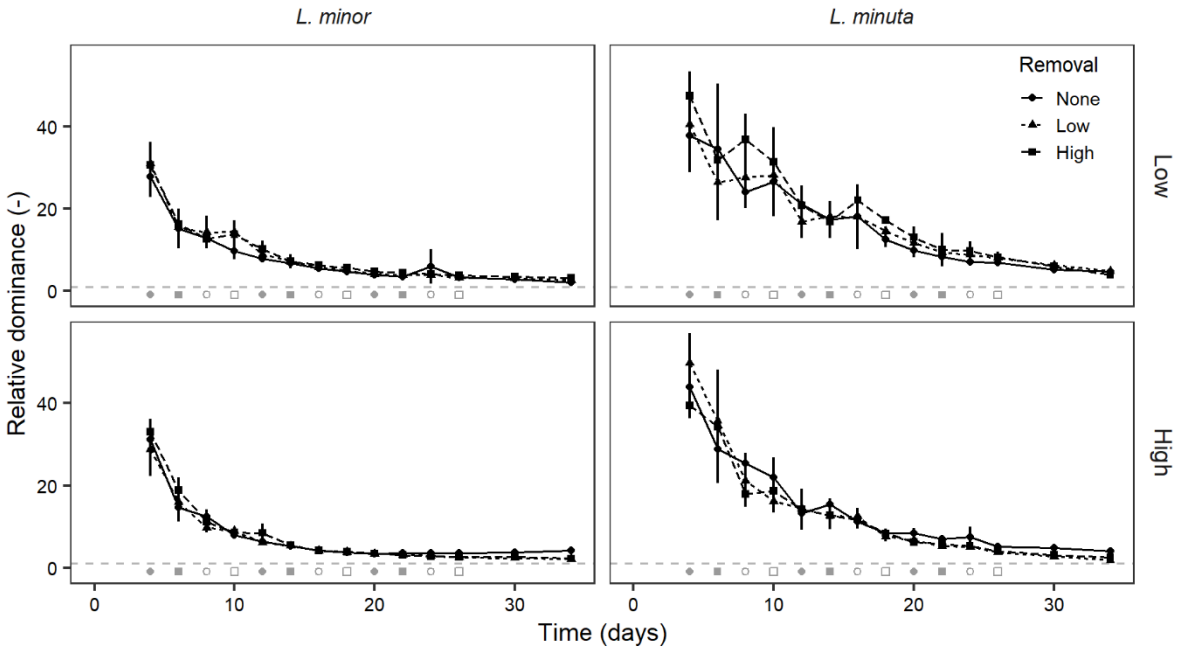

Figure 9.6: Temporal evolution of relative dominance of the primary species. *A decrease in relative dominance is observed due to the continuous increase of the secondary species. Patterns show a lower frequency-dependence when* L. minuta *is introduced, illustrating its higher growth rate. Grey symbols represent introduction (circles) and removal (squares) events, with filled symbols indicating the low frequency pressure.*

9.4 Discussion

9.4.1 Interactions under controlled conditions

Under natural conditions, populations and communities continuously experience external pressures and disturbances, the combined effects of which are hard to predict. To improve understanding, controlled experiments provide relief as they allow to isolate specific pressures and their potential interactions. Here, the individual and interactive effects of two external pressures on the growth and interaction of two duckweed species, *L. minor* and *L. minuta*, were considered, with records of final biomass illustrating the existence of pressure-specific influences. For instance, final biomass of the primary *Lemna* spp. was negatively affected by repetitive partial harvesting, though was hardly affected by the introduction of a secondary species, causing interactive effects to be absent, though dominated by biomass removal if present. In contrast, biomass of the secondary species was positively affected by a higher introduction frequency, especially when combining high-frequency introduction with low-frequency removal. Moreover, secondary species showed to be able to establish a viable population next to the primary species, indicating the absence of a severe negative interaction between *L. minor* and *L. minuta* and confirming reported coexistence (Ceschin *et al.*, 2016; Njambuya *et al.*, 2011).

Growth rates of both *Lemna* spp. varied in time, causing simulations to greatly underestimate the biomass of the primary species by relying on a time-independent growth rate. Temporal patterns of biomass exceeded the simulations due to relatively high initial growth rates, though tended to converge towards the end of the experiment. Moreover, growth rates increased temporarily after each removal occasion, causing differences between scenarios to remain limited. These observations suggest the existence of a scenario-independent endpoint, as the growth rate is negatively affected by biomass density. Under the assumption of such a density-based saturation, external pressures merely affect the time required to reach it. Additionally, it illustrates the negative feedback effect of overcrowding on the growth rate of both *Lemna* spp. at relatively low population size and highlights that density-corrections are crucial when modelling biomass at a temporal level (Driever *et al.*, 2005; Frédéric *et al.*, 2006).

The observed coexistence of both species and existence of a scenario-independent outcome for the primary species are supported by the temporal change in biomass ratio. Over time, the relative dominance of the primary species decreased prior to plateauing above the equilibrium condition. Hence, under mentioned conditions it remains unlikely that the secondary species will assert dominance, affecting subsequent management actions. For instance, biomass removal from a system dominated by *L. minuta* complemented with introduction of *L. minor* will provide a more balanced biomass ratio, but is unlikely to shift towards dominance by *L. minor*.

9.4.2 Interactions under field conditions

Although the performed experiment allows to illustrate certain effects of management on biomass production, caution should be applied during result extrapolation towards field conditions due to a variety of unaddressed factors. For instance, the applied nutrient replenishment can occur within highly dynamic lotic water systems, though represents an unrealistic condition when describing an isolated lentic system. Within both systems, nutrients from the water column are immobilised by growing *Lemna* spp., which is cheered for when treating wastewater in a natural way and under controlled conditions (Muradov *et al.*, 2014; Verma and Suthar, 2014). Yet, when uncontrolled immobilisation causes severe nutrient depletion, opposing *Lemna* spp. might experience a differentiated degree of stress and produce abscisic acid to support the creation of turions (Zhao *et al.*, 2015b). These turions disperse to nearby systems or sink into the sediment, where they remain inactive until better conditions occur. At a larger scale, nutrient immobilisation alters the prevailing biogeochemical cycles, which illustrates the modifying role plants can play within ecosystems (Matsuzaki *et al.*, 2008; Strayer, 2010). This is of special concern when considering alien species, whose invasive success is often linked with their efficiency towards resource use (Paolacci *et al.*, 2016).

Aside from the improved resource use efficiency, a plethora of complementary functional and life-history traits exists to magnify competitive superiority among interacting macrophytes (van Kleunen *et al.*, 2010). For instance, the excretion of allelochemicals degrades habitat suitability by inducing stress and initiating DNA methylation followed by altered gene expression (Zhao *et al.*, 2015b). At a physical level, floating macrophytes have the tendency to create thick mats that impede light penetration within the water column, causing submerged aquatic vegetation to disappear (Driever *et al.*, 2005; Janes *et al.*, 1996). Moreover, it was shown that overcrowding can cause lower growth rates, giving an advantage to faster-growing or more density-tolerant species within these floating mats.

On the other hand, both mutualism and commensalism have been reported between macrophyte species, although being less common for phylogenetically similar species. Both *L. minor* and *L. minuta* showed to be unaffected by the introduction and presence of the opposing species (see Figure 9.3) and thereby confirmed their potential to coexist (Ceschin *et al.*, 2016; Njambuya *et al.*, 2011). However, these inferences are limited to the applied conditions and require additional testing prior to generalisation. This is especially important when aiming to extrapolate the results obtained in this experiment, as the presence of floating separators might have excluded physical interaction processes. For instance, it can be hypothesised that, without barriers, difference in biomass density ($g \cdot m^{-2}$) will affect the observed biomass ratio, with lower density values benefitting the physical overcrowding of the competitor.

9.4.3 Implications for management of invasive alien species

Invasion prevention is agreed upon to be the preferred approach from an economic and ecological perspective (Strayer, 2010; Williams and Grosholz, 2008), though implementation of straightforward guidelines is hampered by ambiguous terminology, international politics and the idiosyncratic behaviour of alien species (Colautti and MacIsaac, 2004; Montgomery *et al.*, 2012). Management of established alien species is traditionally directed towards local eradication and control, while global range shifts induced by climate change are expected to cause more species to disperse and migrate into new territories and to challenge the current definition of being 'alien' (Chen *et al.*, 2011a; Rahel and Olden, 2008). Despite the relatively low success rate, examples of effective eradication within freshwater systems exist and provide a foundation towards future management and tool development for decision-support (Strayer, 2010). Nevertheless, complete eradication remains costly and often highly destructive towards non-target species, which advocates the use of partial, less-destructive eradication programs (Myers *et al.*, 2000). For instance, it was shown that repetitive partial removal of duckweed mats might support the establishment and growth of other macrophyte species by taking advantage of the available physical space.

The obtained results showed that propagule pressure undermines the presence of strong monocultures when partial biomass removal is applied, as indicated by a faster reduction in biomass ratio at high-frequency introduction compared to low-frequency introduction. Hence, the more balanced presence of both *Lemna* spp. due to removal corroborates the effectiveness of partial biomass removal, although the discrepancy between both observations (see Figure 9.6) suggests that *L. minuta* is a slightly stronger competitor (as confirmed by a higher relative growth rate). Invasion by the alien *L. minuta* caused a faster decrease in biomass ratio for the primary *L. minor*, compared to the decrease of *L. minuta* biomass due to introduction of *L. minor*. Consequently, it can be expected that both systems will ultimately reflect a similar state, dominated by *L. minuta*. It remains to be studied how these systems will respond to an additional disturbance event.

Finally, also external factors play a role in steering macrophyte community composition. Within this chapter, the applied removal scenarios assumed the presence of an overall pressure, i.e. the disturbance is not species-specific. Yet, application of species-specific removal (e.g. selective herbivory) can alter final outcomes dramatically (Levine *et al.*, 2004). For instance, selective herbivory of *L. minor* within a system experiencing propagule pressure from *L. minuta* might cause a divergence of the observed temporal trend and ultimately cross the biomass ratio equilibrium faster. Therefore, it is imperative for management to assess the prevailing propagule pressure prior to (partial) biomass harvesting.

9.4.4 Contribution to the study objective

The aim of this chapter was to determine if biomass production is affected by the introduction of a secondary species and whether a different response occurs between native and alien species. Knowledge on the response of a prevailing population facing the introduction of a new species is crucial to develop reactive management plans, allowing case-specific strategies. Moreover, this approach can be extended towards invasive alien species impact reduction by considering the introduced species to be alien. Subsequent establishment depends greatly on the prevailing conditions (both abiotic and biotic), though is often improved via a disturbance (e.g. drought, accidental discharge, harvesting) (Strayer, 2010; Zedler and Kercher, 2005). This disturbance-influenced establishment can be beneficial when a native species is introduced in an alien population, though can be harmful when an alien species is introduced in a native population. Therefore, the response of the prevailing population towards management within the study objective (see Section 1.2.1) was tested with two *Lemna* spp., as these prefer eutrophic conditions and are known to occur as floating mats in ditches, ponds and wetlands (Janse and Van Puijenbroek, 1998). Both species were exposed to (i) biomass removal and (ii) introduction of the opposite species.

Primary production showed to be affected by the performed biomass removal, though was generally unaffected by introduction of a secondary species (see Figure 9.3). More importantly, it showed to be negatively affected by its own growth, as relative growth rates decreased in time (while overall biomass increased). Responses showed a high degree of similarity for *Lemna minor* and *Lemna minuta*, which suggested that no one-way interaction was present and that both species can coexist. These results confirmed reported coexistence in the field (Ceschin *et al.*, 2016; Paolacci *et al.*, 2016) and suggested that biomass removal does not affect the relative abundance of either species. Yet, they additionally insinuated that the native *L. minor* might become less dominant in time due to the introduction of the alien *L. minuta*, compared to *L. minuta* experiencing introduction of *L. minor* (see Figure 9.6). This seemingly minor difference can ultimately result in the suppression of *L. minor* by *L. minuta*, though longer testing conditions are needed to confirm this hypothesis.

The indication that performing partial biomass harvesting within a system exposed to the introduction of a non-established species did not affect the evolution of the relative dominance in time confirmed that both *Lemna* spp. grow relatively independently (Njambuya *et al.*, 2011). This might not be the case for other interactions among macrophyte species, though similar studies are lacking. Nevertheless, it remains recommended to map introduction pressure by neighbouring populations prior to any type of management that causes a temporal disturbance.

9.5 Conclusion

Management of aquatic macrophytes by means of biomass removal provides relief for steering biomass production and community composition. Here, repetitive partial biomass removal delayed the colonisation process and supported higher growth rates for both *Lemna minor* and *Lemna minuta* by reducing the negative feedback due to overcrowding. Similarly, species-specific growth rates decreased in time and showed to be slightly higher (initially and overall) for the alien *L. minuta* compared to the native *L. minor*, corroborating the former's invasive behaviour without being significant. Introduced species were able to establish and coexist with the primary species and benefitted from elevated introduction rates, yet affected the original monoculture differently. More specifically, introduction of *L. minuta* caused a lower relative dominance than the introduction of *L. minor*, potentially due to higher growth rates of the former. Hence, assessment of propagule pressure prior to biomass removal is crucial to avoid the detrimental effects of invasive alien species, making the decision on management intensity and frequency case-dependent. Based on our results, removal of the native *L. minor* to improve light penetration can be performed when pressure by *L. minuta* is absent or low. In contrast, populations of the alien *L. minuta* act as a local species pool and are best removed when *L. minor* is introduced, be it naturally or artificially. As aquatic macrophyte management is a challenging task and will only increase as invasion rates and climate change become more severe, it is imperative to improve understanding of their interacting effects on community composition and ecosystem functioning.

10

General discussion and conclusion

Highlights

- Models and experiments support ecosystem conservation
- Natural dynamics challenge predictions of data-driven models
- Adopting pragmatic approach created caveats and opportunities
- Future perspectives include increased merging of observations and experiments

Abstract

Models and experiments allow to simplify complex natural systems and help understanding patterns and predicting management outcomes. Yet, the majority of ecological research is chopped up in several smaller studies and requires to be comprehensively summarised in order to move from being detailed and confined results to broad and transparent applications. With publicly available data, the influence of data cleaning on model performance was illustrated and concluded on the use of *missForest* to impute missing data and the serial removal of outliers, false absences and redundant variables (both correlated and irrelevant). Threshold values for each pre-processing technique were derived (τ_o = 3, τ_a = 5 %, τ_c = 0.7 and τ_i = 10 %, respectively) and applied prior to inferring macrophyte-specific variable importance scores, which illustrated the importance of and optimal conditions for temperature (> 17 °C), nitrate-N (0.5 mg·L^{-1} up to 1.5 mg·L^{-1}), dissolved oxygen (4 mg·L^{-1} up to 7 mg·L^{-1}), ammonium-N (0.3 mg·L^{-1} up to 0.5 mg·L^{-1}) and pH (7 up to 8.5) to support macrophyte presence. Moreover, model results indicated the potential threat of invasive alien species under prevailing and altered abiotic conditions, although the functional response and relative growth rate did not indicate such a potential under controlled conditions. Integration of the obtained results within wetland management plans provides promising perspectives towards conservation, though identified several areas for future research and improvement. Alternative techniques for data collection, cleaning and analysis are manifold and request testing with respect to applicability and accuracy. Moreover, increased inclusion of functional traits into data-driven models merges the strengths of correlative and process-based modelling, thereby illustrating the inescapable integration of extensive observational data and ecological theory that is essential to tackle the combined threat of climate change and invasive alien species.

10.1 Setting the scene

Throughout previous chapters, the research questions identified in Chapter 1 were systematically tackled, while contributing to the overall study objective (see Section 1.2.1). Challenges were addressed by literature review, correlative modelling and laboratory experiments and contributed to the identification of important abiotic habitat descriptor variables and the value of autecological studies. So far, the results are scattered among the different chapters and require a comprehensive wrap-up, complemented with recommendations for future research.

To start, literature allowed to create an overview of (i) the biotic interactions within shallow freshwater systems, (ii) the obstacles that slow down the implementation of integrated constructed wetlands (ICWs) and (iii) the options for correlative habitat suitability modelling. More specifically, **Chapter 2** summarised the biotic interactions within shallow eutrophic freshwater ecosystems in Table 2.2, while illustrating the capacity of macrophytes to modify the physical and chemical environment into a concert of microhabitats (see Section 2.3.2). In addition, **Chapter 3** compared a selection of correlative modelling techniques for their ease of interpretation, transparency, ecological relevance and predictive performance in order to support technique selection (see Table 3.2). Based on these two chapters, it was decided to focus on (i) macrophytes and (ii) random forests to support wetland management from a biotic perspective.

Secondly, data cleaning and model training allowed the construction of correlative species-specific models. More specifically, data cleaning aimed at improving information density within the provided data and was applied in **Chapter 5** and **Chapter 6**, discussing which imputation technique and which data-specific thresholds to use, respectively. Results showed that *missForest* generally provided the lowest error during imputation (see Section 5.5), while thresholds were selected for combinatory outlier (τ_o = 3), false absence (τ_a = 5 %), correlation (τ_c = 0.7) and irrelevant variable (τ_i = 10 %) identification and elimination (see Table 6.2). Secondly, **Chapter 7** created correlative habitat suitability models (HSMs) that were trained with the pre-processed data and reported that temperature and nitrate highly affected the description of the occupied habitats, being closely followed by ammonium, oxygen and pH (see Figure 7.2). Variable-specific influences allowed to infer general optimal conditions for temperature (> 17 °C), nitrate-N (0.5 mg·L^{-1} up to 1.5 mg·L^{-1}), oxygen (4 mg·L^{-1} up to 7 mg·L^{-1}), ammonium-N (0.3 mg·L^{-1} up to 0.5 mg·L^{-1}) and pH (7 up to 8.5). Based on these conditions, theoretical management scenarios showed to affect habitat suitability in a positive, yet differential, way (see Figure 7.5) and illustrated the need for holistic freshwater management.

Lastly, experiments under controlled conditions are imperative to evaluate the effectiveness of management measures and complement model-based results. Here, experiments were run with the native *Lemna minor* and the alien *Lemna minuta* to (i) identify the applicability of traits to forecast invasive behaviour and (ii) determine the effects of harvesting on biomass ratio. First, **Chapter 8** considered the functional response (resource-based), the relative growth rate (output-based) and a hybrid biomass-based nutrient removal (resource-use efficiency). The observations contradicted the expectations of an invasive alien species being faster in nutrient uptake and biomass production compared to a native species. Secondly, **Chapter 9** considered the potential effects of partial biomass harvesting on overall biomass production. Native-dominated systems showed to be slightly more affected by simultaneous biomass removal and invasion, while alien-dominated systems were characterised by relatively more biomass of the alien species.

Within this chapter, the aim is to frame the individual studies within the overall study objective identified in Chapter 1 (see Section 1.2.1, Section 1.2.2 and Table 10.1). The potential consequences of this work towards wetland conservation are tackled, while specific attention is given to the impending threat of changing environmental conditions and the methodological limitations of the study. Moreover, with the ongoing global changes in mind, future perspectives are identified, prior to concluding this chapter (and the overall work).

Table 10.1: Overview of the individual study objectives as defined in Chapter 1. *For each objective, an internal reference is provided.*

Objective	Topic	Tackled in
1.1	Interacting biotic groups in eutrophic, shallow water bodies	Table 2.2
1.2	Use of habitat modifiers to improve life below water	Section 2.2.2
1.3	Treatment performance to provide clean water and sanitation	Section 2.2.1
1.4	Conclusion on key issues for multifunctional wetlands	Section 2.5
1.5	Overview of advantages and drawbacks of selected techniques	Table 3.2
1.6	Four main steps in ideal modelling procedure	Table 3.1
2.1	Conclusion on comparison of selected imputation techniques	Section 5.5
2.2	Conclusion on threshold values for data pre-processing	Section 6.5
2.3	Performance of species-specific models	Table 7.4
2.4	Variable importance and habitat suitability	Figure 7.1 & 7.2
2.5	Identification of potential prevalence and management effects	Figure 7.4 & 7.5
3.1	Defining calculation of trait indices	Section 8.2.3
3.2	Individual traits versus ecosystem-based techniques	Section 8.4.4
3.3	Temporal evolution of biomass and biomass ratio	Figure 9.4 & 9.6
3.4	Biomass of two *Lemna* spp. under different treatments	Figure 9.3

10.2 Contribution to the conservation of wetlands

Wetland conservation entails three main groups of management activities: protection, restoration and construction (see Box 1.2). Each of these groups benefits from the development of habitat suitability models as illustrated by their application to delineate reserve areas (Elith *et al.*, 2006; Real *et al.*, 2006), guide restoration efforts (Keshtkar *et al.*, 2013; Van der Lee *et al.*, 2006), predict distributions of native and alien species (Boets *et al.*, 2013; Chefaoui and Lobo, 2008) and explore the potential effects of climate change (Barbet-Massin *et al.*, 2014; Domisch *et al.*, 2013). To extend these observations towards macrophyte-based freshwater management, correlative habitat suitability models were developed in Chapter 7. More specifically, conditional random forests were trained for 58 different macrophyte species due to their de-correlated ensemble-based approach and reported outperformance of more conventional modelling techniques (Benito *et al.*, 2013; Breiman, 2001; Guo *et al.*, 2015; Strobl *et al.*, 2007).

The main contribution of the HSMs developed in Chapter 7 towards wetland conservation (and, by extension, general freshwater management) is the identification of macrophyte-specific response curves. Based on these curves, two main types of management approaches can be distinguished: (i) the prevailing conditions are considered to be fixed boundary conditions or (ii) the prevailing conditions are flexible and can be adapted to optimally support a specific (set of) macrophyte(s). Both approaches entail some degree of biotic control (e.g. harvesting, eradication, manual introduction), though only the latter considers additional abiotic control (e.g. intensive pre-treatment, chemical precipitation). More importantly, the resulting management plans can be extended by including the response curves of alien species during decision-making. This is illustrated in Box 10.1, which represents the potential implementation of the models developed in Chapter 7.

Yet, HSMs are limited in the answers they can provide to support the development of management plans, especially when dealing with questions related to (i) rare (e.g. endangered, alien) species, (ii) biotic interactions, (iii) dispersal dynamics or (iv) occurrence probability (Araújo *et al.*, 2005; Bruneel *et al.*, 2018; Gallien *et al.*, 2010). Responses to these issues associated with single-species abiotic HSMs include the use of multilayer models (Dubuis *et al.*, 2011; Guisan and Rahbek, 2011), the inclusion of biotic predictors (Giannini *et al.*, 2013), the integration of remote sensing (Cord *et al.*, 2014) and the implementation of model calibration (Jarnevich *et al.*, 2015). Alternatively, experiments under controlled conditions are performed to provide a clearer causal link between an explanatory and response variable in comparison to the correlations extracted by HSMs. However, such experiments often produce results that are only valid in particular environmental settings, which limits their extrapolation potential and overall ecological relevance (Fagúndez and Lema, 2019; Forbes *et al.*, 2008).

Box 10.1: Example of model contribution to freshwater management

The models developed in Chapter 7 provided information on (i) variable importance towards delineating the occupied range, (ii) species-specific and overall response curves, (iii) temporal potential prevalence patterns and (iv) the value of case-specific management. For instance, under the assumed nutrient enrichment in the considered wetland configuration (see Section 1.2.1), *Phragmites australis* and *Lemna minor* depict a similar habitat suitability index (HSI) score (see Figure 7.5; NUT scenario in 2010), being higher than the remaining three species. However, if the presence of *Ceratophyllum demersum* is preferred over *P. australis* and *L. minor*, management can aim at avoiding the establishment of the latter two species (e.g. by eradicating prevailing populations), while no specific additional actions towards abiotic conditions is performed, providing *C. demersum* with the highest HSI score (i.e. the NUT-BAU scenario). Establishment of the latter can occur naturally (e.g. originating from neighbouring species pools) or artificially (e.g. manual introduction), though remains conditional to the abiotic habitat environment.

Simultaneously, similar information can be retrieved for guiding alien species management, including (i) the preferred abiotic conditions, (ii) the potential geographical distribution and (iii) the impact of management on HSI scores. For instance, given similar nutrient enrichment (see previous paragraph), HSI scores for the alien *Lemna minuta* are lower compared to *P. australis*, *L. minor* and *C. demersum*, though increase in time. Actual survival of these species remains conditional to the abiotic environment, though can result in an increasing level of competition between *L. minuta* and *C. demersum* in time, especially when the establishment of *P. australis* and *L. minor* is artificially avoided (see above). However, due to contrasting growth forms of the floating *L. minuta* and the submerged *C. demersum*, it is expected that the former will outcompete the latter.

The main contribution of the experiments performed in Chapter 8 and Chapter 9 towards wetland conservation (and, by extension, general freshwater management) is the framework used to assess the invasion potential of an alien species. Based on this framework, information is gathered to support the delineation of (i) proactive and (ii) reactive management plans of alien species. By means of comparative trait-based assessment (e.g. nutrient use, growth rate, stress tolerance), an alien species can be classified as less, equally or more invasive or impactful than a (co-generic) species, which helps in prioritising alien species management (Early *et al.*, 2016). This is illustrated in Box 10.2, which represents the added value of the experiments performed in Chapter 8 and Chapter 9.

Box 10.2: Example of experiment contribution to freshwater management

The models developed in Chapter 7 allowed the comparison of preferred abiotic conditions for the native *Lemna minor* and alien *L. minuta*. The resulting species response curves indicated a relatively similar correlation of temperature, nitrate and oxygen with habitat suitability, showing overall higher habitat suitability index (HSI) scores for the native *L. minor* compared to the alien *L. minuta* (Figure 7.2). These observations were corroborated by most locations favouring the presence of the native *L. minor* during scenario analysis (Figure 7.5) and within the Limnodata Neerlandica (Figure 7.6). Still, some sites tended to be more suitable for *L. minuta*, while discrepancies in HSI scores for both species are expected to decrease further due to increasing temperatures. The resulting effect on the survival and establishment of the alien *L. minuta* after introduction in a site with reported presence of *L. minor* cannot be inferred from these models and requires (i) a more process-based approach or (ii) experiments under controlled conditions to derive (i) the invasive behaviour of a species and (ii) the potential impact on existing population(s).

The experiments performed in Chapter 8 showed that differences occurred in nutrient uptake (Figure 8.1 and Figure 8.2) and biomass production (Figure 8.4) between *L. minor* and *L. minuta*. More specifically, *L. minor* took up more nutrients and created more dry biomass than *L. minuta*, which suggests that the latter is less invasive than (and potentially relatively similar to) the native *L. minor*. Yet, it also suggests that the prevailing nutrient dynamics and the associated ecosystem functioning are likely to change if a transition from a native-based to an alien-based system (e.g. due to extreme propagule pressure and higher suitability scores) occurs. Overall, the experiments did not confirm the invasive behaviour of the alien *L. minuta* and thereby advise against the universal use of the applied traits to forecast the invasive behaviour of new alien species.

In addition, the experiments performed in Chapter 9 illustrated that the introduction and survival of the alien *L. minuta* did not affect the biomass production of the native *L. minor* and vice versa. Moreover, even under increasing partial harvesting stress, biomass production of *L. minor* remained largely unaffected by the introduction of *L. minuta* and vice versa (Figure 9.1). Yet, it showed that relatively higher biomass ratios were obtained for the invasive *L. minuta* due to slightly higher growth rates compared to the native *L. minor* (Figure 9.4). Hence, management can aim at reducing the introduction of the alien *L. minuta* to maintain higher dominance by the native *L. minor*, though considering the limited impact and the assumed functional similarity, priority can be assigned to more harmful alien species.

10.2.1 Changing environments

Models and experiments allow to simplify complex natural systems, help understanding patterns and predict management outcomes. Yet, conditions continuously change due to endogenic and exogenic processes and pressures, which challenge model transferability and experiment relevance. These processes occur on local (e.g. plant-based nutrient uptake, settling of suspended solids), regional (e.g. habitat creation, micro-climates) and global (e.g. climate change) scales, thereby changing community composition and functioning. Each of these changes at the abiotic level has the potential to disrupt vulnerable communities and cause local disappearance of one (or more) species, thereby reflecting the inherently idiosyncratic behaviour of natural systems.

Changes and disturbances at the abiotic level are expected to extend beyond the individual level and alter complete ecological networks by affecting resource availability and interaction intensity (Davis *et al.*, 2000; Walther, 2010). The inherent interaction displayed by each organism with its environment, alters both the abiotic habitat conditions and the resulting community composition in both space and time (Vitousek, 1990). For instance, the use of macrophytes to mitigate elevated pollutant levels by means of phytoremediation (see Box 2.2), supports better conditions for other species to grow and underlies many restoration projects relying on natural succession. Similarly, regional changes in land use have caused better land drainage (e.g. urbanisation) and increased fertiliser use (e.g. agriculture), thereby negatively affecting downstream processes in river basins with peak flows and eutrophication, respectively (Kingsford *et al.*, 2016). At a larger scale, climate change is expected to affect hydrological patterns and temperatures, causing higher disturbance frequencies and magnitudes to occur and weaken established communities (IPCC, 2014). Therefore, predictions of future species distributions under altered abiotic conditions need to be considered with care as species respond differently to changes and violate the assumption of niche conservation (Dormann *et al.*, 2012). Adaptability to rapidly changing conditions by altering phenology, physiology or morphology is therefore highlighted as a main trait for providing species with a competitive advantage. Generally, high levels of plasticity and adaptability are characteristic for many invasive species (Davidson *et al.*, 2011), though predictions of their distribution are frequently underestimations due to violating the equilibrium assumption (Gallien *et al.*, 2012).

The results obtained throughout this work remain valuable under changing environmental conditions as they indicate species-specific preferred environmental conditions (**Chapter 7**) and illustrate trait-specific differences among physiologically and phylogenetically similar species (**Chapter 8** and **Chapter 9**). By taking these results into account, management should be able to (i) focus on key habitat descriptors, (ii) focus on key species, (iii) infer invasion potential differently and (iv) define harvesting strategies.

10.2.2 Limitations of the study

Aside from the contributions outlined in previous sections, a variety of limitations and caveats were identified throughout this study. The delineation of the working field performed in Section 1.2.1 aided in narrowing down the scope of the individual chapters towards data cleaning, model development and experimental design. Simultaneously, it created several areas of caution, including (i) the data being used, (ii) the techniques being selected and (iii) the experimental design being applied.

10.2.2.1Data used

The characteristics and content of the Limnodata Neerlandica were outlined in Chapter 4 and clearly indicate various potential points of criticism. First of all, the data combines information from a variety of institutes that have been performing field assessments for multiple years, without applying a single standardised methodology. Consequently, data collection was highly institute- and campaign-specific and resulted in high levels of missing data (see Figure 4.1). Moreover, the database does not include an overview of the methodologies, protocols and equipment used by the institutes to collect physicochemical data, which requires the assumption that all values for a single variable were recorded in a similar manner (regardless of institute and sampling campaign). The inclusion of metadata remains a common challenge in data-driven analyses.

Secondly, macrophyte occurrence was recorded with a variety of techniques (see Appendix, Table A.2) and contained several undefined and hybrid species. These techniques tend to vary in the spatial extent covered during assessment, ranging from small quadrants to (relatively) large stretches. Due to this variety, the discretisation into a presence/absence-statement can be considered as too simplistic. Moreover, misidentifications might occur, causing both false presences and false absences to be included in the data. Hence, the use of this macrophyte data to evaluate and assess water quality within the Netherlands is not recommended (Verdonschot and van Oosten-Siedlecka, 2010). Similarly, correlative analyses are expected to be negatively affected by these issues, though it was assumed that these effects remained relatively limited.

Thirdly, the majority of the macrophyte species were characterised by low levels of prevalence (see Figure 4.4C), including rare, endangered and recently-introduced alien species. An arbitrary cut-off value of 100 presences (i.e. 200 observations in a balanced data set) was assumed to provide sufficient information and to limit the overall number of macrophyte species. Lower numbers have been reported in literature (e.g. 135 (Guo *et al.*, 2015), 120 (Forio *et al.*, 2015), 110 (Vezza *et al.*, 2015)), with 30 observations being considered the minimum (Jarnevich *et al.*, 2015). Additional backing of the cut-off value was provided after performing data reduction and maintaining only 20 variables, which allows providing roughly 10 instances per variable. Due to this approach, the selected macrophytes are relatively generalist species, while excluding most specialist species.

Lastly, imputation of missing data was performed on the physicochemical data within the common data set (see Section 4.2.3.2), after a reduction in the degree of missing data (i.e. from 93.7 % to 49.7 %; Section 6.2.1). More specifically, imputation was based on the associations between the explanatory variables in the common data, which is merely a subset of the available physicochemical data (i.e. not all physicochemical data were linked with a biotic response variable). This indicates inefficient use of the available information, though reflects a higher similarity with most occurrence-based correlative modelling studies. Nevertheless, variable associations derived from the complete physicochemical data set might be able to provide more accurate estimates of the missing data points and definitely merits further study.

10.2.2.2 Technique selection

The development of data-driven habitat suitability models relies on two main components: (1) the quality of the collected data and (2) technique selection (Segurado and Araújo, 2004). Data cleaning has a positive effect on the quality of the data and the associated model results (Kotsiantis *et al.*, 2006; Maldonado *et al.*, 2015), although the actual impact differs among the various techniques that are available. Similarly, a plethora of modelling techniques exists to correlate species occurrence with environmental conditions, without a single-best approach being identified (Jarnevich *et al.*, 2015; Lawson *et al.*, 2014). Most studies apply subjective technique selection based on previous experience or recommendations from literature, while a more case-specific comparative approach provides a higher potential to improve model accuracy. Still, these comparisons are biased by the selection of techniques being included, as performed in Chapter 3 (modelling techniques), Chapter 5 (imputation techniques) and Chapter 6 (pre-processing techniques).

The selection of modelling techniques was narrowed down to commonly used data-driven techniques that were able to deal with presence-absence data (PA; see Section 3.1). As such, several presence-only (PO) modelling techniques were excluded from the comparison, including environmental envelopes (e.g. BIOCLIM, HABITAT) (Tsoar *et al.*, 2007), ecological niche factor analysis (ENFA) (Hirzel *et al.*, 2002), maximum entropy (MAXENT) (VanDerWal *et al.*, 2009) and point-process models (Renner *et al.*, 2015). Especially the use of point-process models is noteworthy due to their possibility to fit spatial and temporal patterns, while interpretation and implementation are relatively straightforward (Renner *et al.*, 2015). The technique is highly linked with MAXENT (Aarts *et al.*, 2012; Renner and Warton, 2013), though has only been limitedly applied in ecology due to its relatively recent introduction. Nevertheless, the availability of PA data within this study supported the applied delineation of the chapter, along with reports on PA-based models outperforming PO models (Brotons *et al.*, 2004; Elith *et al.*, 2006; Phillips *et al.*, 2009).

Similarly, the selection of imputation techniques was narrowed down to obtain a selection that (i) provided a single data set as output and (2) was able to deal with the *missing at random* (MAR) mechanism. These criteria excluded various imputation techniques, including multiple-value (e.g. multivariate normal imputation (Lee and Carlin, 2010) and multiple imputation via chained equation (Schmitt *et al.*, 2015)) and techniques able to deal with the *not missing at random* (NMAR) mechanism. The latter requires a more advanced statistical approach than the techniques dealing with the MAR mechanisms, which limits their availability in commonly available software packages for data analysis. For instance, Liu *et al.* (2018) designed an information decomposition imputation (IDIM) algorithm using fuzzy memberships to deal with missing data, illustrating its case-specificity. Aside from these criteria-based exclusions, a plethora of single-value imputation techniques were arbitrarily omitted, including Bayesian principal component analysis (Oba *et al.*, 2003), singular value decomposition (Alter *et al.*, 2000), fuzzy k-means (Li *et al.*, 2004) and artificial neural networks (Chandramouli *et al.*, 2007).

Lastly, technique selection occurred to narrow the options for data pre-processing towards mostly statistical techniques and the associated threshold(s). Alternative approaches range along the objective-subjective continuum for outliers (e.g. percentile-based exclusion, expert-based assessment, visual inspection (Gobeyn *et al.*, 2017)), correlated (e.g. expert-based (Sauer *et al.*, 2011)) and irrelevant (e.g. iterative model development (Gregorutti *et al.*, 2017), expert-based (Brandt *et al.*, 2017)) variable removal. Identification and elimination of false absences is rarely reported despite the awareness on their negative impact on model performance (Gu and Swihart, 2004; Lobo *et al.*, 2010). Yet, the potential of including false absences in the training data often restricts the selection of background or pseudo-absence data (Chefaoui and Lobo, 2008; Phillips *et al.*, 2009).

10.2.2.3Experimental design

Experiments under controlled conditions provide crucial information on causal processes, biotic interactions and treatment effects. The design of the experiments in this work entailed a series of choices that can be considered arbitrary and open to discussion, including (i) the selection of *Lemna* spp. as test species, (ii) the applied test conditions and (iii) the selection of traits. To start, the alien *Lemna minuta* and the native *Lemna minor* were selected based on (1) the assumed eutrophic conditions (see Section 1.2.1), (2) their widespread occurrence within Europe (Hussner, 2012) and (3) the existence of guidelines for testing conditions (see also Section 4.3). Moreover, their high reproduction rate and manipulability added a pragmatic basis for selecting *Lemna* spp. (Ceschin *et al.*, 2016; Njambuya *et al.*, 2011; Paolacci *et al.*, 2016; Paolacci *et al.*, 2018). Similar tests can be performed with the alien *Acorus calamus* and *Elodea nuttallii* to complement the developed models (Table C.2).

Secondly, test conditions were selected based on the guidelines for performing ecotoxicological test with *L. minor* (OECD, 2006). Light intensity, temperature and composition of the growth medium were defined according to these guidelines, which limits the ecological relevance of the experiments and the associated extrapolation capacity of the results (Fagúndez and Lema, 2019; Forbes *et al.*, 2008). For instance, the lowest concentration in the trait-based experiment for total phosphorus was 1.33 ± 0.01 $mg \cdot L^{-1}$, while Flemish waters contained on average about 0.48 $mg \cdot L^{-1}$ in 2018 (VMM, 2019). In contrast, the lowest total nitrogen concentration was 4.2 ± 0.1 $mg \cdot L^{-1}$ and was highly similar to the concentration in Flemish surface waters (i.e. about 4.5 $mg \cdot L^{-1}$) (VMM, 2019). It remains possible that different results will be obtained when applying more ecologically relevant testing conditions.

Thirdly, resource use and biomass growth were considered as traits, because of their simplicity and relevance towards invasion and outcompetition. Yet, a variety of alternative traits exists, including specific leaf area (SLA), leaf thickness, leaf nutrient concentration, light-saturated photosynthetic rate and dark respiration (Pérez-Harguindeguy *et al.*, 2013). Each of these traits can contribute partially to an overall competitive advantage, although their relative contribution can be altered by limiting the phylogenetic differences (Strauss *et al.*, 2006; van Kleunen *et al.*, 2010). More importantly, only single values for each trait were inferred, while many species are characterised by a certain degree of trait plasticity. Species containing higher trait plasticity are considered (i) to be more tolerant towards stressors (ii) to have a competitive advantage over other species and (iii) to have a steering effect on community dynamics (Barbour *et al.*, 2019; Bellavance and Brisson, 2010; Berg and Ellers, 2010). Hence, increased trait plasticity is often hypothesised to positively affect the invasive success of alien species (Berg and Ellers, 2010; Davidson *et al.*, 2011).

10.2.2.4Performance interpretation versus real data

A recurrent issue in environmental data science is the evaluation of the applied techniques. Observations and results are often treated in an objective (or statistical) manner and represented by a single (set of) metric(s), e.g. outlier removal based on the threshold $\tau_0 = 3$, imputation accuracy assessment with the normalised root mean squared error (NRMSE) and model performance evaluation using the area under the receiver operating characteristic curve (AUC). Aside from simplifying understanding, comparability and repeatability, no information on the ecological validity is included in these thresholds or metrics. More specifically, various valid questions remain, including (i) *Is imputation really accurate and what are the differences with actual data?* (ii) *Are outliers, false absences, correlations and irrelevant variables correctly (i.e. ecologically-founded) removed?* and (iii) *Do predicted presences correspond with observed presences and are there patterns in the misidentifications?* Such legitimate questions remain difficult to answer when dealing with relatively large amounts of data.

10.3 Future perspectives

The research outlined throughout this work responds to increased needs of efficient land use, nature development, mitigation of climate change, improved circular economy and, above all, fighting the ongoing biodiversity loss within freshwater systems (Harrison *et al.*, 2018; He *et al.*, 2019). The steps taken throughout this work are small in comparison to the spatial and temporal dimensions of these problems, but contribute to governing a framework that can provide answers to the challenges faced by society. It would be presumptuous to state that this work was the final hurdle to be taken, as various improvements and extensions are waiting to be implemented and investigated. Within the following sections, specific attention is given to potential and promising advances related to model development and invasive alien species management, framed around the consequences of environmental change as the proverbial elephant in the room.

10.3.1 Model development

The application of data-driven modelling techniques experienced a rapid increase due to unprecedented growths of publicly available data and technological progress in computational capacity. However, these reasons may well be the main drawbacks of data-driven modelling and warrants careful application. More specifically, data extracted from publicly available databases have a tendency to be incomplete, dirty and of generally low quality, especially when the data originates from various contributors (Hernández and Stolfo, 1998; Maldonado *et al.*, 2015). Within this work, detailed data cleaning identified unique space-time combinations of abiotic conditions and macrophyte observations and included a comparison of several imputation and pre-processing techniques. Yet, despite aiming for a practical procedure that allows application in other studies, several alternative approaches, methodologies and recommendations have been excluded from this work due to pragmatic reasons. The subsequent sections shortly introduce these alternatives and additionally identify topics for future research and exploration.

10.3.1.1 Data availability and collection

An element of major importance with respect to the data used for observation-based modelling is the inclusion of both metadata and relevant explanatory variables (Austin and Van Niel, 2011; Barbet-Massin *et al.*, 2014; Braunisch *et al.*, 2013). Technological improvements steer data collection forward by supporting non-destructive sampling campaigns and high-resolution data (both temporal and spatial). For instance, spectral reflectance of leaves can be used to determine the degree of stress experienced by plants without having to analyse leaf content biochemically (Fagúndez and Lema, 2019). At a larger scale, remote sensing has shown to improve model quality by including more local variables within a correlative model, thereby supplementing standard field data collection with valuable explanatory variables (Bruneel *et al.*, 2018; Cord *et al.*, 2014).

Aside from improving predictor selection, promising results have already been obtained with the sampling and analysis of environmental DNA (eDNA) to characterise the prevailing community (Bohmann *et al.*, 2014). By relying on eDNA, there is no need to visually confirm species presence, while the chance of false absences and false presences decreases. Hence, the detection rate of rare, endangered and invasive species is positively affected.

10.3.1.2Data cleaning

From **Chapter 5**, it was derived that the ensemble-based *missForest* algorithm performs better than the other selected techniques within the provided context. Combining *missForest* with *k*-nearest neighbours and iterative least square regression to construct an ensemble of imputation techniques (thus a multiple-value imputation) was considered to be outside the scope of the comparison, yet merits further scrutiny. Moreover, imputation uncertainty within the final data sets (i.e. in Chapter 7) was assumed to be low due to the size of the data, yet no formal analysis was performed. Hence, future studies should focus on (i) the potential of ensemble imputation, (ii) the discrepancy between single-value and multiple-value imputation and (iii) the corroboration of the results from this work with alternative data sets.

Secondly, the identified thresholds from **Chapter 6** related to subsequent data pre-processing with respect to outliers, false absences and redundant variables can be used as a guideline for future data-driven model development. However, the identification of outliers and false absences required the selection of a method-specific threshold α, which was arbitrarily fixed and expected to additionally affect the final number of instances. Therefore, future research on data pre-processing can entail (i) how the choice of α during outlier or false absence elimination affects data availability and model performance, (ii) how alternative pre-processing techniques affect data set characteristics and model performance, (iii) how the order of pre-processing techniques changes model performance and (iv) how the type of data influences threshold selection and values.

10.3.1.3Habitat Suitability Models

The lack of a single-best modelling technique renders selection into a subjective procedure. More specifically, selection is influenced by literature reporting unequivocal results when comparing techniques, which underlines the advantages of ensemble-based modelling (Araújo and New, 2007; Araújo *et al.*, 2005b; Austin, 2007; Svetnik *et al.*, 2003). Within this work, the choice for random forests within **Chapter 7** to link species occurrence with abiotic conditions was invoked by the fact that the ensemble approach increases model stability and decreases overfitting (Breiman, 2001; Strobl *et al.*, 2007). Yet, these advantages come at the expense of transparency and computation time.

Ensembles of and errors in correlative models

The obtained response from these random forests merely reflected species-specific habitat suitability, without providing a statement on the predicted probability of species occurrence. Thus, additional care is needed to infer species distributions or, at a higher level, species richness and community composition (Dubuis *et al.*, 2011). For instance, the inclusion of dispersal dynamics allows further fine-tuning of the results, while the use of a logistic curve or a fixed threshold provides a statistical approach to obtain a continuous or binary statement on species-specific occurrence probability, respectively. Based on these probabilities, species richness and community composition can be derived by stacking multiple species-specific models (S-SDMs), though results are prone to be overly positive due to the exclusion of ecological assembly rules (Dubuis *et al.*, 2011; Guisan and Rahbek, 2011). Dubuis *et al.* (2011) suggested to counter this overprediction by curtailing the community by means of a single macro-ecological model (MEM), developed to predict species richness. By combining both approaches, an accurate estimation of species richness is obtained (MEM) and supplemented with the expected species to be present (S-SDM).

An important point of attention with respect to correlative modelling is the inherent error propagation and the resulting uncertainty (Guisan and Zimmerman, 2000). Although having been partly reduced by the progress in statistical modelling, errors are introduced due to statistical limitations and confined understanding of the biological systems (Elith *et al.*, 2006; Fielding and Bell, 1997). Reduction of the uncertainty related to biotic interactions can be achieved by a variety of actions, including (i) continuity of basic biological and ecological research to account for biotic interactions, (ii) the systematic collection of species occurrence, (iii) the monitoring over time to validate existing models and (iv) the creation of awareness of overall uncertainty (Braunisch *et al.*, 2013; Elith and Leathwick, 2009; Sinclair *et al.*, 2010). Furthermore, algorithm improvement and climate scenarios have been the main focus in literature dealing with error introduction, thereby unfairly neglecting the importance of predictor selection (Barbet-Massin *et al.*, 2014). Consequently, important ecological drivers might be missed, causing linkages between ecological theory and model configuration to be weak or even non-existing (Austin, 2002; Elith and Leathwick, 2009).

Data-driven versus process-based models

One of the hailed and most criticised characteristics of data-driven habitat suitability and species distribution models (SDMs) is their potential to predict future species distributions (Austin and Van Niel, 2011; Braunisch *et al.*, 2013; Guisan *et al.*, 2006). Purely data-driven models (e.g. decision trees, GLMs, ANNs) are developed based on observational data without substantial integration of existing ecological knowledge. Therefore, they only describe the current situation (i.e. the realised niche) and are more or less limited to the range of the observed predictor values (Dormann *et al.*, 2012).

Considering that future environmental conditions can lead to predictor values situated outside this range, indicates that purely data-driven models might not be the best option for predicting future species distributions (Braunisch *et al.*, 2013; Dormann *et al.*, 2012). Furthermore, species prevalence is not only determined by abiotic characteristics and currently existing situations, but also by the ability of dispersion, the carrying capacity of the environment and the possibility of competitive exclusion due to co-occurrences (Austin and Van Niel, 2011; Guisan and Rahbek, 2011). These aspects are not easily included in a purely data-driven model structure. As a result, not all SDMs are optimally suited for predicting future species distributions in light of climate change.

On the other hand, models that combine data and knowledge (e.g. fuzzy logic, BBNs) provide the ability to extend the range of predictor values beyond the observed range and to include ecological interactions (e.g. dispersion rate, carrying capacity, competition). By combining data and a certain degree of knowledge, models can shift from being data-driven to become more process-based, thereby supporting the prediction of future species distributions with a more ecologically sound basis (Dormann *et al.*, 2012). In short, future model development will have to focus more on combining observational data, ecological theories and expert knowledge rather than being purely data-driven, in order to increase the reliability of model-based species distribution predictions.

Nevertheless, the added value of data-driven habitat suitability models towards management should not be underestimated, as climate change, habitat destruction and invasive species are continuously shaping new environments. Species experiencing these altered abiotic conditions are forced to adapt or migrate, causing shifts in distribution patterns and unprecedented extinction rates (Chen *et al.*, 2011a; Rahel and Olden, 2008; Walther, 2010). Moreover, due to these high rates of global change, abiotic conditions might change faster than the dispersal rate of macrophytes and cause local extinctions of native populations (Bornette and Puijalon, 2011). Hence, attempting restoration via manually introducing native species might turn out to be futile when the prevailing conditions do not support species presence, which illustrates and highlights the potential of habitat suitability models. Simultaneously, geographic range shifts of nearby populations provide an opportunity to maintain ecological functioning and structure, though challenges the definition of what constitutes an alien species and, consequently, conservation management in general (Rahel and Olden, 2008).

10.3.2 Managing invasive alien species

Performing autecological experiments under controlled conditions provides valuable information on species-specific functional traits, population dynamics, biotic interactions and disturbance resilience (Hofstra *et al.*, 2020). Functional traits have been applied frequently to infer competitive dominance between co-existing species, though showed in **Chapter 8** that under the considered conditions no statement could be inferred on the invasive behaviour of the alien *Lemna minuta*. In contrast, reactive management showed to be an interesting approach to reduce *Lemna* spp. dominance in **Chapter 9**, though requires a preliminary assessment of propagule pressure prior to deployment.

Both experiments add to the existing knowledge on invasive alien species management, yet indicate that additional testing is needed to derive species-wide, condition-independent trait values and field-relevant management scenarios. Extension towards other light regimes, temperature patterns, nutritional conditions, stressor combinations and biotic communities is therefore highly recommended, especially when predictions beyond the observed environmental conditions are requested to infer the consequences of ecosystem disturbance (see also Box 10.3) (Fagúndez and Lema, 2019).

Early-succession traits (e.g. minimal temperature for seed emergence, ratio of photosynthetic tissues) are crucial in steering species survival and determining competitive outcomes, although require simultaneous assessment of dispersal traits to quantify actual propagule pressures. Hence, it is expected that no single trait provides a clear, unequivocal statement on invasive behaviour and that multi-trait evaluation is needed to categorise alien species (van Kleunen *et al.*, 2010; Zedler and Kercher, 2004). This highlights that any contribution to the species-specific trait database is to be supported, even when no spectacular results are obtained.

Aside from the alterations in abiotic conditions causing range shifts and reduced resistance to invasion, also ecological interactions are affected by climate change, with invasive alien species potentially profiting from it (e.g. change in parasitism, diseases, competitors and predators) (Walther *et al.*, 2009). The relatively limited attention towards freshwater systems along with the complex interaction between climate change and invasion impact governs the development of new and unfamiliar ecosystems, requiring adaptive management in uncharted fields (Rahel and Olden, 2008; Strayer, 2010). For instance, Kelly *et al.* (2015) showed that replacement of the native *Elodea canadensis* by the alien *E. nuttallii* in Irish lakes hardly affected physicochemical conditions or biomass production, yet the significant differences in oxygen levels and plant community composition illustrated the structural and functional change caused by invasion.

Box 10.3: Temporal characteristics of disturbances

Establishment of alien species within ecosystems largely depends on the occurrence of disturbances, the effects of which are conditional to both frequency and intensity (Bornette and Puijalon, 2011; Catford *et al.*, 2009; Strayer, 2010). For instance, combined sewer overflows during peak precipitation introduce additional nutrients into the receiving water body, but temporal effects remain limited due to natural dilution processes. In contrast, construction of hydropower infrastructure affects hydrologic conditions up- and downstream for an indefinite amount of time. The reduced biotic resistance resulting from these events provides an optimal opportunity for new colonisers to take advantage and establish viable populations, with various alien species among them (Davis *et al.*, 2000; Zedler and Kercher, 2004).

Long-term consequences of these changes are unknown and hard to predict due to their dynamic nature, but include the facilitated introduction of non-indigenous species by established alien species, causing an invasional meltdown (Montgomery *et al.*, 2012; Simberloff and Von Holle, 1999; Williams and Grosholz, 2008). Empirical evidence of this hypothesis is limited and underlines the necessity for further fundamental and applied research to counter the indecisiveness in management, the proliferation of hypotheses and the study bias towards terrestrial and marine invasions (Montgomery *et al.*, 2012; Rahel and Olden, 2008; Simberloff, 2006). By studying the invasion process, unique information is gathered on biotic interactions and overall ecosystem functioning, allowing the identification of attention points during management projects (Myers *et al.*, 2000; Strayer, 2010; Williams and Grosholz, 2008).

By predicting future distributions of alien species, species distribution models (SDMs) provide potential to be used in risk assessment by forecasting the effect of future alien species distributions on native species (e.g. Gallardo *et al.* (2012), Kolar and Lodge (2002), Reichard and Hamilton (1997)). For instance, Gallardo and Aldridge (2013) investigated the combined threat of climate change and invasive alien species on native species and reported that, based on SDM predictions, native species will experience considerable losses. Furthermore, they observed that climate change does not necessarily influence invasive alien species distribution in a positive way. However, due to uncertainties related to adaptation potential, SDMs might even underestimate the future spread of invasive alien species (Gallardo and Aldridge, 2013), thereby underlining the necessity for additional biological and ecological research. More specifically, experimental studies that extend knowledge on functional traits allow to parameterise process-based models in an ecologically relevant way and thereby provide a sound basis for extrapolating predictions outside the currently occurring environmental domain (Dormann *et al.*, 2012).

10.4 Concluding remarks

The application of habitat suitability and species distribution models within ecosystem management is rapidly increasing due to the relentless rise of data dimensionality and augmented awareness on the innumerable services provided by ecosystems (Dormann *et al.*, 2012). However, progress has occurred mostly at the level of algorithm development and thereby largely ignored other sources of uncertainty, including ecological theory, data cleaning, predictor selection and model transferability (Barbet-Massin *et al.*, 2014). Within data-driven models, ecological relevance is crucial to distinguish between finding effective ecological relationships and pure pattern extraction, thereby supporting the acceptance and applicability of otherwise black-box models (Austin and Van Niel, 2011; Brewer *et al.*, 2016).

Considering the current rates of changes occurring at all spatial scales, it is expected that data-driven modelling will increase as ecosystem managers and decision-makers are more often looking towards science for answers. It is believed that this work positively contributes to future studies by discussing data cleaning techniques, model applications and controlled experiments. More specifically, it was found that *missForest* can accurately impute missing data, while the identification of outliers, false absences, correlated and important variables helps developing ecologically relevant models by applying specific thresholds (τ_o = 3, τ_a = 5 %, τ_c = 0.7 and τ_i = 10 %, respectively). Also, baseline hyperparameter settings for random forests (*ntree* = 200 and 10 repetitions) were identified along with optimal environmental conditions for the five most important habitat descriptors (temperature > 17 °C, nitrate-N = 0.5 mg·L^{-1} up to 1.5 mg·L^{-1}, oxygen = 4 mg·L^{-1} up to 7 mg·L^{-1}, ammonium-N = 0.3 mg·L^{-1} up to 0.5 mg·L^{-1} and pH = 7 up to 8.5). Lastly, controlled experiments provided information on key traits and interactions and created a basis for future research on alternative conditions and interactions.

Still, numerous challenges are identified related to the rise of data-driven modelling and the consequent translation of results into policies. Global biodiversity informatics progresses but continues to face several hurdles, including non-digitised collections, limited knowledge sharing and overall isolation. More cooperation in a world that contains more biogeographers outside the historical developed regions and a focus on the more biodiverse tropical regions remains a main goal to fight the ongoing biodiversity crisis (Peterson *et al.*, 2015) and to comply with the Aichi targets within the Strategic Plan for Biodiversity 2011-2020 (CBD, 2020). Moreover, the translation of scientific results and recommendations into policies often works as a retardant towards conservation, being additionally exacerbated by the sheer extent of the affected areas. Within this context, it should remain clear that models can help decision-making, while remaining a mere simplification of reality. Or, as stated by Box and Draper (1987): "*All models are wrong, but some are useful*".

References

Aarts, G.; Fieberg, J. and Matthiopoulos, J. (2012) Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution* **3** (1), 177-187, doi: 10.1111/j.2041-210X.2011.00141.x.

Acevedo, P.; Jiménez-Valverde, A.; Lobo, J. M. and Real, R. (2012) Delimiting the geographical background in species distribution modelling. *Journal of Biogeography* **39** (8), 1383-1390, doi: 10.1111/j.1365-2699.2012.02713.x.

Adriaenssens, V.; Baets, B. D.; Goethals, P. L. M. and Pauw, N. D. (2004a) Fuzzy rule-based models for decision support in ecosystem management. *Science of the Total Environment* **319** (1–3), 1-12, doi: 10.1016/S0048-9697(03)00433-9.

Adriaenssens, V.; Goethals, P. L. M.; Charles, J. and De Pauw, N. (2004b) Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. *Annales de Limnologie - International Journal of Limnology* **40** (03), 181-191, doi: 10.1051/limn/2004016.

Aguilera, P. A.; Fernández, A.; Fernández, R.; Rumí, R. and Salmerón, A. (2011) Bayesian networks in environmental modelling. *Environmental Modelling & Software* **26** (12), 1376-1388, doi: 10.1016/j.envsoft.2011.06.004.

Aguilera, P. A.; Fernández, A.; Reche, F. and Rumí, R. (2010) Hybrid Bayesian network classifiers: Application to species distribution models. *Environmental Modelling & Software* **25** (12), 1630-1639, doi: 10.1016/j.envsoft.2010.04.016.

Ahmadi-Nedushan, B.; St-Hilaire, A.; Bérubé, M.; Robichaud, É.; Thiémonge, N. and Bobée, B. (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications* **22** (5), 503-523, doi: 10.1002/rra.918.

Alexander, M. E.; Dick, J. T. A.; Weyl, O. L. F.; Robinson, T. B. and Richardson, D. M. (2014) Existing and emerging high impact invasive species are characterized by higher functional responses than natives. *Biology Letters* **10** (2), doi: 10.1098/rsbl.2013.0946.

Alford, R. A. and Richards, S. J. (1999) Global Amphibian Declines: A Problem in Applied Ecology. *Annual Review of Ecology and Systematics* **30** (1), 133-165, doi: 10.1146/annurev.ecolsys.30.1.133.

Allouche, O.; Tsoar, A. and Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43** (6), 1223-1232, doi: 10.1111/j.1365-2664.2006.01214.x.

Alpert, P.; Bone, E. and Holzapfel, C. (2000) Invasiveness, invasibility and the role of environmental stress in the spread of non-native plants. *Perspectives in Plant Ecology, Evolution and Systematics* **3** (1), 52-66, doi: 10.1078/1433-8319-00004.

Alter, O.; Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97** (18), 10101, doi: 10.1073/pnas.97.18.10101.

Ambelu, A.; Mekonen, S.; Koch, M.; Addis, T.; Boets, P.; Everaert, G. and Goethals, P. (2014) The Application of Predictive Modelling for Determining Bio-Environmental Factors Affecting the Distribution of Blackflies (Diptera: Simuliidae) in the Gilgel Gibe Watershed in Southwest Ethiopia. *PLoS ONE* **9** (11), e112221, doi: 10.1371/journal.pone.0112221.

Amon, J. P.; Agrawal, A.; Shelley, M. L.; Opperman, B. C.; Enright, M. P.; Clemmer, N. D.; Slusser, T.; Lach, J.; Sobolewski, T.; Gruner, W. and Entingh, A. C. (2007) Development of a

wetland constructed for the treatment of groundwater contaminated by chlorinated ethenes. *Ecological Engineering* **30** (1), 51-66, doi: 10.1016/j.ecoleng.2007.01.008.

Anderson, R. P. and Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography* **37** (7), 1378-1393, doi: 10.1111/j.1365-2699.2010.02290.x.

Angélibert, S.; Marty, P.; Céréghino, R. and Giani, N. (2004) Seasonal variations in the physical and chemical characteristics of ponds: implications for biodiversity conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems* **14** (5), 439-456, doi: 10.1002/aqc.616.

Appenroth, K.-J.; Sree, K. S.; Böhm, V.; Hammann, S.; Vetter, W.; Leiterer, M. and Jahreis, G. (2017) Nutritional value of duckweeds (Lemnaceae) as human food. *Food Chemistry* **217**, 266-273, doi: 10.1016/j.foodchem.2016.08.116.

Araújo, M. B. and Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33** (10), 1677-1688, doi: 10.1111/j.1365-2699.2006.01584.x.

Araújo, M. B. and New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* **22** (1), 42-47, doi: 10.1016/j.tree.2006.09.010.

Araújo, M. B.; Pearson, R. G.; Thuiller, W. and Erhard, M. (2005a) Validation of species–climate impact models under climate change. *Global Change Biology* **11** (9), 1504-1513, doi: 10.1111/j.1365-2486.2005.01000.x.

Araújo, M. B.; Whittaker, R. J.; Ladle, R. J. and Erhard, M. (2005b) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography* **14** (6), 529-538, doi: 10.1111/j.1466-822X.2005.00182.x.

Archer, K. J. and Kimes, R. V. (2008) Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* **52** (4), 2249-2260, doi: 10.1016/j.csda.2007.08.015.

Armstrong, R. A. (2014) When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* **34** (5), 502-508, doi: 10.1111/opo.12131.

Aronson, J.; Floret, C.; Le Floc'h, E.; Ovalle, C. and Pontanier, R. (1993) Restoration and Rehabilitation of Degraded Ecosystems in Arid and Semi-Arid Lands. I. A View from the South. *Restoration Ecology* **1** (1), 8-17, doi: 10.1111/j.1526-100X.1993.tb00004.x.

Arthur, E. L.; Rice, P. J.; Rice, P. J.; Anderson, T. A.; Baladi, S. M.; Henderson, K. L. D. and Coats, J. R. (2005) Phytoremediation—An Overview. *Critical Reviews in Plant Sciences* **24** (2), 109-122, doi: 10.1080/07352680590952496.

Assilian, S. *Artificial intelligence in control of real dynamic systems* PhD thesis, London, (1974).

Austin, M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling* **200** (1), 1-19, doi: 10.1016/j.ecolmodel.2006.07.005.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157** (2–3), 101-118, doi: 10.1016/S0304-3800(02)00205-3.

Austin, M. P. and Van Niel, K. P. (2011) Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography* **38** (1), 1-8, doi: 10.1111/j.1365-2699.2010.02416.x.

Auvinen, H.; Du Laing, G.; Meers, E. and Rousseau, D. P. L. (2016) Constructed Wetlands Treating Municipal and Agricultural Wastewater – An Overview for Flanders, Belgium in *Natural and Constructed Wetlands: Nutrients, heavy metals and energy cycling, and flow* (ed J. Vymazal) 179-207 (Springer International Publishing, Cham, Switzerland).

Bakker, E. S.; Sarneel, J. M.; Gulati, R. D.; Liu, Z. and van Donk, E. (2013) Restoring macrophyte diversity in shallow temperate lakes: biotic versus abiotic constraints. *Hydrobiologia* **710** (1), 23-37, doi: 10.1007/s10750-012-1142-9.

Balcombe, C. K.; Anderson, J. T.; Fortney, R. H. and Kordek, W. S. (2005a) Aquatic macroinvertebrate assemblages in mitigated and natural wetlands. *Hydrobiologia* **541** (1), 175-188, doi: 10.1007/s10750-004-5706-1.

Balcombe, C. K.; Anderson, J. T.; Fortney, R. H. and Kordek, W. S. (2005b) Vegetation, Invertebrate, and Wildlife Community Rankings and Habitat Analysis of Mitigation Wetlands in West Virginia. *Wetlands Ecology and Management* **13** (5), 517-530, doi: 10.1007/s11273-004-5074-7.

Baldy, V.; Thiebaut, G.; Fernandez, C.; Sagova-Mareckova, M.; Korboulewsky, N.; Monnier, Y.; Perez, T. and Tremolieres, M. (2015) Experimental Assessment of the Water Quality Influence on the Phosphorus Uptake of an Invasive Aquatic Plant: Biological Responses throughout Its Phenological Stage. *PLoS ONE* **10** (3), e0118844, doi: 10.1371/journal.pone.0118844.

Banks-Leite, C. and Ewers, R. M. (2009) Ecosystem Boundaries. *eLS*, doi: 10.1002/9780470015902.a0021232.

Barbet-Massin, M.; Jetz, W. and Heikkinen, R. (2014) A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Diversity and Distributions* **20** (11), 1285-1295, doi: 10.1111/ddi.12229.

Barbour, M. A.; Erlandson, S.; Peay, K.; Locke, B.; Jules, E. S. and Crutsinger, G. M. (2019) Trait plasticity is more important than genetic variation in determining species richness of associated communities. *Journal of Ecology* **107** (1), 350-360, doi: 10.1111/1365-2745.13014.

Barde, M. P. and Barde, P. J. (2012) What to use to express the variability of data: Standard deviation or standard error of mean? *Perspectives in clinical research* **3** (3), 113-116, doi: 10.4103/2229-3485.100662.

Barker, T.; Hatton, K.; O'Connor, M.; Connor, L. and Moss, B. (2008) Effects of nitrate load on submerged plant biomass and species richness: results of a mesocosm experiment. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* **173** (2), 89-100, doi: 10.1127/1863-9135/2008/0173-0089.

Barnaby, W. (2009) Do nations go to war over water? *Nature* **458** (7236), 282-283, doi: 10.1038/458282a.

Barrat-Segretain, M.-H. (2005) Competition between Invasive and Indigenous Species: Impact of Spatial Pattern and Developmental Stage. *Plant Ecology* **180** (2), 153-160, doi: 10.1007/s11258-004-7374-7.

Basheer, I. A. and Hajmeer, M. (2000) Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods* **43** (1), 3-31, doi: 10.1016/S0167-7012(00)00201-3.

Batzer, D. P.; Pusateri, C. R. and Vetter, R. (2000) Impact of Fish Predation on Marsh Invertebrates: Direct and Indirect Effects. *Wetlands* **20** (2), 307-312, doi: 10.1672/0277-5212(2000)020[0307:IOFPOM]2.0.CO;2.

Batzer, D. P. and Resh, V. H. (1991) Trophic Interactions among a Beetle Predator, a Chironomid Grazer, and Periphyton in a Seasonal Wetland. *Oikos* **60** (2), 251-257, doi: 10.2307/3544872.

Beaver, J. R.; Miller-Lemke, A. M. and Acton, J. K. (1998) Midsummer zooplankton assemblages in four types of wetlands in the Upper Midwest, USA. *Hydrobiologia* **380** (1), 209-220, doi: 10.1023/A:1003452118351.

Becerra-Jurado, G.; Harrington, R. and Kelly-Quinn, M. (2012) A review of the potential of surface flow constructed wetlands to enhance macroinvertebrate diversity in agricultural landscapes with particular reference to Integrated Constructed Wetlands (ICWs). *Hydrobiologia* **692** (1), 121-130, doi: 10.1007/s10750-011-0866-2.

Becerra Jurado, G.; Callanan, M.; Gioria, M.; Baars, J. R.; Harrington, R. and Kelly-Quinn, M. (2009) Comparison of macroinvertebrate community structure and driving environmental factors in natural and wastewater treatment ponds. *Hydrobiologia* **634** (1), 153-165, doi: 10.1007/s10750-009-9900-z.

Bellavance, M.-E. and Brisson, J. (2010) Spatial dynamics and morphological plasticity of common reed (Phragmites australis) and cattails (Typha sp.) in freshwater marshes and roadside ditches. *Aquatic Botany* **93** (2), 129-134, doi: 10.1016/j.aquabot.2010.04.003.

Benito, B. M.; Cayuela, L. and Albuquerque, F. S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution* **4** (4), 327-335, doi: 10.1111/2041-210x.12022.

Bennetsen, E.; Gobeyn, S. and Goethals, P. L. M. (2016) Species distribution models grounded in ecological theory for decision support in river management. *Ecological Modelling* **325** (Supplement C), 1-12, doi: 10.1016/j.ecolmodel.2015.12.016.

Bennett, N. D.; Croke, B. F. W.; Guariso, G.; Guillaume, J. H. A.; Hamilton, S. H.; Jakeman, A. J.; Marsili-Libelli, S.; Newham, L. T. H.; Norton, J. P.; Perrin, C.; Pierce, S. A.; Robson, B.; Seppelt, R.; Voinov, A. A.; Fath, B. D. and Andreassian, V. (2013) Characterising performance of environmental models. *Environmental Modelling & Software* **40**, 1-20, doi: 10.1016/j.envsoft.2012.09.011.

Benyamine, M.; Bäckström, M. and Sandén, P. (2004) Multi-Objective Environmental Management in Constructed Wetlands. *Environmental Monitoring and Assessment* **90** (1), 171-185, doi: 10.1023/B:EMAS.0000003577.22824.8e.

Berg, M. P. and Ellers, J. (2010) Trait plasticity in species interactions: a driving force of community dynamics. *Evolutionary Ecology* **24** (3), 617-629, doi: 10.1007/s10682-009-9347-8.

Berg, P.; Moseley, C. and Haerter, J. O. (2013) Strong increase in convective precipitation in response to higher temperatures. *Nature Geoscience* **6** (3), 181-185, doi: 10.1038/ngeo1731.

Bergstra, J. and Bengio, Y. (2012) Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13** (Feb), 281-305.

Bergström, I.; Mäkelä, S.; Kankaala, P. and Kortelainen, P. (2007) Methane efflux from littoral vegetation stands of southern boreal lakes: An upscaled regional estimate. *Atmospheric Environment* **41** (2), 339-351, doi: 10.1016/j.atmosenv.2006.08.014.

Beutel, M. W.; Morgan, M. R.; Erlenmeyer, J. J. and Brouillard, E. S. (2014) Phosphorus Removal in a Surface-Flow Constructed Wetland Treating Agricultural Runoff. *Journal of Environmental Quality* **43** (3), 1071-1080, doi: 10.2134/jeq2013.11.0463.

Blaas, H. and Kroeze, C. (2016) Excessive nitrogen and phosphorus in European rivers: 2000–2050. *Ecological Indicators* **67**, 328-337, doi: 10.1016/j.ecolind.2016.03.004.

Blackburn, T. M.; Pyšek, P.; Bacher, S.; Carlton, J. T.; Duncan, R. P.; Jarošík, V.; Wilson, J. R. U. and Richardson, D. M. (2011) A proposed unified framework for biological invasions. *Trends in Ecology & Evolution* **26** (7), 333-339, doi: 10.1016/j.tree.2011.03.023.

Bø, T. H.; Dysvik, B. and Jonassen, I. (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research* **32** (3), e34-e34, doi: 10.1093/nar/gnh026.

Boets, P.; Holguin, G. J. E.; Lock, K. and Goethals, P. L. M. (2013a) Data-driven habitat analysis of the Ponto-Caspian amphipod Dikerogammarus villosus in two invaded regions in Europe. *Ecological Informatics* **17** (0), 36-45, doi: 10.1016/j.ecoinf.2012.07.001.

Boets, P.; Lock, K. and Goethals, P. L. M. (2013b) Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. *Ecological Informatics* **17** (0), 73-81, doi: 10.1016/j.ecoinf.2012.06.001.

Boets, P.; Lock, K.; Messiaen, M. and Goethals, P. L. M. (2010) Combining data-driven methods and lab studies to analyse the ecology of Dikerogammarus villosus. *Ecological Informatics* **5** (2), 133-139, doi: 10.1016/j.ecoinf.2009.12.005.

Boets, P.; Michels, E.; Meers, E.; Lock, K.; Tack, F. M. G. and Goethals, P. L. M. (2011) Integrated Constructed Wetlands (ICW): Ecological Development in Constructed Wetlands for Manure Treatment. *Wetlands* **31** (4), 763-771, doi: 10.1007/s13157-011-0193-4.

Bohmann, K.; Evans, A.; Gilbert, M. T. P.; Carvalho, G. R.; Creer, S.; Knapp, M.; Yu, D. W. and de Bruyn, M. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution* **29** (6), 358-367, doi: 10.1016/j.tree.2014.04.003.

Bolton, L.; Joseph, S.; Greenway, M.; Donne, S.; Munroe, P. and Marjo, C. E. (2019) Phosphorus adsorption onto an enriched biochar substrate in constructed wetlands treating wastewater. *Ecological Engineering: X* **1**, 100005, doi: 10.1016/j.ecoena.2019.100005.

Boomer, I.; Aladin, N.; Plotnikov, I. and Whatley, R. (2000) The palaeolimnology of the Aral Sea: a review. *Quaternary Science Reviews* **19** (13), 1259-1278, doi: 10.1016/S0277-3791(00)00002-0.

Born, W.; Rauschmayer, F. and Bräuer, I. (2005) Economic evaluation of biological invasions—a survey. *Ecological Economics* **55** (3), 321-336, doi: 10.1016/j.ecolecon.2005.08.014.

Bornette, G. and Puijalon, S. (2011) Response of aquatic plants to abiotic factors: a review. *Aquatic Sciences* **73** (1), 1-14, doi: 10.1007/s00027-010-0162-7.

Box, G. E. P. and Draper, N. R. (1987) *Empirical model-building and response surfaces*. Vol. 424 (Wiley New York.).

Brain, R. A. and Cedergreen, N. (2008) Biomarkers in aquatic plants: selection and utility in *Reviews of environmental contamination and toxicology* 49-109 (Springer).

Bramm, M. E.; Lassen, M. K.; Liboriussen, L.; Richardson, K.; Ventura, M. and Jeppesen, E. (2009) The role of light for fish–zooplankton–phytoplankton interactions during winter in shallow lakes – a climate change perspective. *Freshwater Biology* **54** (5), 1093-1109, doi: 10.1111/j.1365-2427.2008.02156.x.

Brandt, L. A.; Benscoter, A. M.; Harvey, R.; Speroterra, C.; Bucklin, D.; Romañach, S. S.; Watling, J. I. and Mazzotti, F. J. (2017) Comparison of climate envelope models developed using expert-selected variables versus statistical selection. *Ecological Modelling* **345**, 10-20, doi: 10.1016/j.ecolmodel.2016.11.016.

Braunisch, V.; Coppes, J.; Arlettaz, R.; Suchant, R.; Schmid, H. and Bollmann, K. (2013) Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography* **36** (9), 971-983, doi: 10.1111/j.1600-0587.2013.00138.x.

Breiman, L. (2001) Random Forests. *Machine Learning* **45** (1), 5-32, doi: 10.1023/A:1010933404324.

Breitburg, D.; Levin, L. A.; Oschlies, A.; Grégoire, M.; Chavez, F. P.; Conley, D. J.; Garçon, V.; Gilbert, D.; Gutiérrez, D.; Isensee, K.; Jacinto, G. S.; Limburg, K. E.; Montes, I.; Naqvi, S. W. A.; Pitcher, G. C.; Rabalais, N. N.; Roman, M. R.; Rose, K. A.; Seibel, B. A.; Telszewski, M.; Yasuhara, M. and Zhang, J. (2018) Declining oxygen in the global ocean and coastal waters. *Science* **359** (6371), doi: 10.1126/science.aam7240.

Brewer, M. J.; O'Hara, R. B.; Anderson, B. J. and Ohlemüller, R. (2016) Plateau: a new method for ecologically plausible climate envelopes for species distribution modelling. *Methods in Ecology and Evolution* **7** (12), 1489-1502, doi: 10.1111/2041-210X.12609.

Brey, T.; Jarre-Teichmann, A. and Borlich, O. (1996) Artificial neural network versus multiple linear regression prediciting P/B ratios from empirical data. *Marine ecology-progress series* **140**, 251-256.

Brisson, J. and Chazarenc, F. (2009) Maximizing pollutant removal in constructed wetlands: should we pay more attention to macrophyte species selection? *Science of the Total Environment* **407** (13), 3923-3930, doi: 10.1016/j.scitotenv.2008.05.047.

Brix, H. (1997) Do Macrophytes Play a Role in Constructed Treatment Wetlands? *Water Science Technology* **35** (5), 11-17, doi: 10.1016/S0273-1223(97)00047-4.

Brock, G. N.; Shaffer, J. R.; Blakesley, R. E.; Lotz, M. J. and Tseng, G. C. (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics* **9** (1), 12, doi: 10.1186/1471-2105-9-12.

Bronstein, J. L. (1994) Our Current Understanding of Mutualism. *The Quarterly Review of Biology* **69** (1), 31-51, doi: 10.1086/418432.

Brooker, R. W.; Maestre, F. T.; Callaway, R. M.; Lortie, C. L.; Cavieres, L. A.; Kunstler, G.; Liancourt, P.; Tielbörger, K.; Travis, J. M. J.; Anthelme, F.; Armas, C.; Coll, L.; Corcket, E.; Delzon, S.; Forey, E.; Kikvidze, Z.; Olofsson, J.; Pugnaire, F.; Quiroz, C. L.; Saccone, P.; Schiffers, K.; Seifan, M.; Touzard, B. and Michalet, R. (2008) Facilitation in plant communities: the past, the present, and the future. *Journal of Ecology* **96** (1), 18-34, doi: 10.1111/j.1365-2745.2007.01295.x.

Brosse, S.; Guegan, J.-F.; Tourenq, J.-N. and Lek, S. (1999) The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecological Modelling* **120** (2–3), 299-311, doi: 10.1016/S0304-3800(99)00110-6.

Brotons, L.; Thuiller, W.; Araújo, M. B. and Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27** (4), 437-448, doi: 10.1111/j.0906-7590.2004.03764.x.

Bruneel, S.; Gobeyn, S.; Verhelst, P.; Reubens, J.; Moens, T. and Goethals, P. (2018) Implications of movement for species distribution models - Rethinking environmental data tools. *Science of the Total Environment* **628-629**, 893-905, doi: 10.1016/j.scitotenv.2018.02.026.

Burks, R. L.; Jeppesen, E. and Lodge, D. M. (2000) Macrophyte and fish chemicals suppress Daphnia growth and alter life-history traits. *Oikos* **88** (1), 139-147, doi: 10.1034/j.1600-0706.2000.880116.x.

Calero, S.; Segura, M.; Rojo, C. and Rodrigo, M. A. (2015) Shifts in plankton assemblages promoted by free water surface constructed wetlands and their implications in eutrophication remediation. *Ecological Engineering* **74** (Supplement C), 385-393, doi: 10.1016/j.ecoleng.2014.11.003.

Callaway, R. M. and Walker, L. R. (1997) Competition and Facilitation: A Synthetic Approach to Interactions in Plant Communities. *Ecology* **78** (7), 1958-1965, doi: 10.1890/0012-9658(1997)078[1958:CAFASA]2.0.CO;2.

Cao, L.; Guisen, D.; Bingbin, H.; Qingyi, M.; Huimin, L.; Zijian, W. and Fu, S. (2007) Biodiversity and water quality variations in constructed wetland of Yongding River system. *Acta Ecologica Sinica (International Journal)* **27** (9), 3670-3677, doi: 10.1016/S1872-2032(07)60080-8.

Caraco, N. F. and Cole, J. J. (2002) Contrasting impacts of a native and alien macrophyte on dissolved oxygen in a large river. *Ecological Applications* **12** (5), 1496-1509, doi: 10.1890/1051-0761(2002)012[1496:CIOANA]2.0.CO;2.

Carlsson, N. O. L. and Brönmark, C. (2006) Size-dependent effects of an invasive herbivorous snail (Pomacea canaliculata) on macrophytes and periphyton in Asian wetlands. *Freshwater Biology* **51** (4), 695-704, doi: 10.1111/j.1365-2427.2006.01523.x.

Carr, G. M.; Duthie, H. C. and Taylor, W. D. (1997) Models of aquatic plant productivity: a review of the factors that influence growth. *Aquatic Botany* **59** (3), 195-215, doi: 10.1016/S0304-3770(97)00071-5.

Carvalho, P. N.; Basto, M. C. P.; Almeida, C. M. R. and Brix, H. (2014) A review of plant–pharmaceutical interactions: from uptake and effects in crop plants to phytoremediation in constructed wetlands. *Environmental Science and Pollution Research* **21** (20), 11729-11763, doi: 10.1007/s11356-014-2550-3.

Castelletti, A. and Soncini-Sessa, R. (2007) Bayesian Networks and participatory modelling in water resource management. *Environmental Modelling & Software* **22** (8), 1075-1088, doi: 10.1016/j.envsoft.2006.06.003.

Castro-Castellon, A. T.; Chipps, M. J.; Hankins, N. P. and Hughes, J. M. R. (2016) Lessons from the "Living-Filter": An in-reservoir floating treatment wetland for phytoplankton reduction prior to a water treatment works intake. *Ecological Engineering* **95** (Supplement C), 839-851, doi: 10.1016/j.ecoleng.2016.07.023.

Catalano, A. S.; Lyons-White, J.; Mills, M. M. and Knight, A. T. (2019) Learning from published project failures in conservation. *Biological Conservation* **238**, 108223, doi: 10.1016/j.biocon.2019.108223.

Catford, J. A.; Jansson, R. and Nilsson, C. (2009) Reducing redundancy in invasion ecology by integrating hypotheses into a single theoretical framework. *Diversity and Distributions* **15** (1), 22-40, doi: 10.1111/j.1472-4642.2008.00521.x.

Cavicchioli, R.; Ripple, W. J.; Timmis, K. N.; Azam, F.; Bakken, L. R.; Baylis, M.; Behrenfeld, M. J.; Boetius, A.; Boyd, P. W.; Classen, A. T.; Crowther, T. W.; Danovaro, R.; Foreman, C. M.; Huisman, J.; Hutchins, D. A.; Jansson, J. K.; Karl, D. M.; Koskella, B.; Mark Welch, D. B.; Martiny, J. B. H.; Moran, M. A.; Orphan, V. J.; Reay, D. S.; Remais, J. V.; Rich, V. I.; Singh, B. K.; Stein, L. Y.; Stewart, F. J.; Sullivan, M. B.; van Oppen, M. J. H.; Weaver, S. C.; Webb, E. A. and Webster, N. S. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology* **17** (9), 569-586, doi: 10.1038/s41579-019-0222-5.

CBD. (2020) *Convention on Biological Diversity*, <https://www.cbd.int/> (Last accessed on 30/01/2020).

Cedergreen, N. and Madsen, T. V. (2002) Nitrogen uptake by the floating macrophyte *Lemna minor*. *New Phytologist* **155** (2), 285-292, doi: 10.1046/j.1469-8137.2002.00463.x.

Celton, M.; Malpertuy, A.; Lelandais, G. and de Brevern, A. G. (2010) Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics* **11** (1), 15, doi: 10.1186/1471-2164-11-15.

Céréghino, R.; Ruggiero, A.; Marty, P. and Angélibert, S. (2008) Biodiversity and distribution patterns of freshwater invertebrates in farm ponds of a south-western French agricultural landscape. *Hydrobiologia* **597** (1), 43-51, doi: 10.1007/s10750-007-9219-6.

Ceschin, S.; Abati, S.; Leacche, I.; Iamonico, D.; Iberite, M. and Zuccarello, V. (2016) Does the alien Lemna minuta show an invasive behavior outside its original range? Evidence of antagonism with the native L. minor in central Italy. *International Review of Hydrobiology* **101** (5-6), 173-181, doi: 10.1002/iroh.201601841.

Chandramouli, V.; Brion, G.; Neelakantan, T. R. and Lingireddy, S. (2007) Backfilling missing microbial concentrations in a riverine database using artificial neural networks. *Water Research* **41** (1), 217-227, doi: 10.1016/j.watres.2006.08.022.

Chawaka, S. N.; Boets, P.; Goethals, P. L. M. and Mereta, S. T. (2018) Does the protection status of wetlands safeguard diversity of macroinvertebrates and birds in southwestern Ethiopia? *Biological Conservation* **226**, 63-71, doi: 10.1016/j.biocon.2018.07.021.

Chefaoui, R. M. and Lobo, J. M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* **210** (4), 478-486, doi: 10.1016/j.ecolmodel.2007.08.010.

Chen, I. C.; Hill, J. K.; Ohlemüller, R.; Roy, D. B. and Thomas, C. D. (2011a) Rapid Range Shifts of Species Associated with High Levels of Climate Warming. *Science* **333** (6045), 1024, doi: 10.1126/science.1206432.

Chen, P.-Y.; Lee, P.-F.; Ko, C.-J.; Ko, C.-H.; Chou, T.-C. and Teng, C.-J. (2011b) Associations Between Water Quality Parameters and Planktonic Communities in Three Constructed Wetlands, Taipei. *Wetlands* **31** (6), 1241-1248, doi: 10.1007/s13157-011-0236-x.

Choi, J.-Y.; Jeong, K.-S.; Kim, S.-K.; La, G.-H.; Chang, K.-H. and Joo, G.-J. (2014) Role of macrophytes as microhabitats for zooplankton community in lentic freshwater ecosystems of South Korea. *Ecological Informatics* **24**, 177-185, doi: 10.1016/j.ecoinf.2014.09.002.

Colautti, R. I. and MacIsaac, H. J. (2004) A neutral terminology to define 'invasive' species. *Diversity and Distributions* **10** (2), 135-141, doi: 10.1111/j.1366-9516.2004.00061.x.

Colin, N.; Porte, C.; Fernandes, D.; Barata, C.; Padrós, F.; Carrassón, M.; Monroy, M.; Cano-Rocabayera, O.; de Sostoa, A.; Piña, B. and Maceda-Veiga, A. (2016) Ecological relevance of biomarkers in monitoring studies of macro-invertebrates and fish in Mediterranean rivers. *Science of the Total Environment* **540**, 307-323, doi: 10.1016/j.scitotenv.2015.06.099.

Comín, F. A.; Romero, J. A.; Hernández, O. and Menéndez, M. (2001) Restoration of Wetlands from Abandoned Rice Fields for Nutrient Removal, and Biological Community and Landscape Diversity. *Restoration Ecology* **9** (2), 201-208, doi: 10.1046/j.1526-100x.2001.009002201.x.

Cord, A. F.; Klein, D.; Gernandt, D. S.; de la Rosa, J. A. P. and Dech, S. (2014) Remote sensing data can improve predictions of species richness by stacked species distribution models: a case study for Mexican pines. *Journal of Biogeography* **41** (4), 736-748, doi: 10.1111/jbi.12225.

Costanza, R.; de Groot, R.; Sutton, P.; van der Ploeg, S.; Anderson, S. J.; Kubiszewski, I.; Farber, S. and Turner, R. K. (2014) Changes in the global value of ecosystem services. *Global Environmental Change* **26**, 152-158, doi: 10.1016/j.gloenvcha.2014.04.002.

Cronk, J. K. and Mitsch, W. J. (1994) Periphyton productivity on artificial and natural surfaces in constructed freshwater wetlands under different hydrologic regimes. *Aquatic Botany* **48** (3), 325-341, doi: 10.1016/0304-3770(94)90024-8.

Culley, D. D. and Epps, E. A. (1973) Use of Duckweed for Waste Treatment and Animal Feed. *Journal Water Pollution Control Federation* **45** (2), 337-347.

Cutler, D. R.; Edwards, T. C.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J. and Lawler, J. J. (2007) Random Forests for Classification in Ecology. *Ecology* **88** (11), 2783-2792, doi: 10.1890/07-0539.1.

D'Heygere, T.; Goethals, P. L. M. and De Pauw, N. (2003) Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecological Modelling* **160** (3), 291-300, doi: 10.1016/S0304-3800(02)00260-0.

Dale, H. M. (1986) Temperature and light: The determining factors in maximum depth distribution of aquatic macrophytes in Ontario, Canada. *Hydrobiologia* **133** (1), 73-77, doi: 10.1007/BF00010804.

Davic, R. D. and Welsh, H. H. (2004) On the Ecological Roles of Salamanders. *Annual Review of Ecology, Evolution, and Systematics* **35** (1), 405-434, doi: 10.1146/annurev.ecolsys.35.112202.130116.

Davidson, A. M.; Jennions, M. and Nicotra, A. B. (2011) Do invasive species show higher phenotypic plasticity than native species and, if so, is it adaptive? A meta-analysis. *Ecology Letters* **14** (4), 419-431, doi: 10.1111/j.1461-0248.2011.01596.x.

Davidson, N. C. (2014) How much wetland has the world lost? Long-term and recent trends in global wetland area. *Marine and Freshwater Research* **65** (10), 934-941, doi: 10.1071/MF14173.

Davis, M. A.; Grime, J. P. and Thompson, K. (2000) Fluctuating resources in plant communities: a general theory of invasibility. *Journal of Ecology* **88** (3), 528-534, doi: 10.1046/j.1365-2745.2000.00473.x.

Davis, M. A. and Thompson, K. (2000) Eight Ways to Be a Colonizer; Two Ways to Be an Invader: A Proposed Nomenclature Scheme for Invasion Ecology. *Bulletin of the Ecological Society of America* **81** (3), 226-230.

Davison, A. C. (2001) Biometrika Centenary: Theory and General Methodology. *Biometrika* **88** (1), 13-52, doi: 10.2307/2673674.

De'ath, G. and Fabricius, K. E. (2000) Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis. *Ecology* **81** (11), 3178-3192, doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

De Troyer, N.; Mereta, T. S.; Goethals, L. P. and Boets, P. (2016) Water Quality Assessment of Streams and Wetlands in a Fast Growing East African City. *Water* **8** (4), doi: 10.3390/w8040123.

Dedecker, A. P.; Goethals, P. L. M.; Gabriels, W. and De Pauw, N. (2004) Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling* **174** (1–2), 161-173, doi: 10.1016/j.ecolmodel.2004.01.003.

Demars, B. O. L. and Edwards, A. C. (2009) Distribution of aquatic macrophytes in contrasting river systems: A critique of compositional-based assessment of water quality. *Science of the Total Environment* **407** (2), 975-990, doi: 10.1016/j.scitotenv.2008.09.012.

Dhir, B.; Sharmila, P. and Saradhi, P. P. (2009) Potential of Aquatic Macrophytes for Removing Contaminants from the Environment. *Critical Reviews in Environmental Science and Technology* **39** (9), 754-781, doi: 10.1080/10643380801977776.

Dick, J. T. A.; Alexander, M. E.; Jeschke, J. M.; Ricciardi, A.; MacIsaac, H. J.; Robinson, T. B.; Kumschick, S.; Weyl, O. L. F.; Dunn, A. M.; Hatcher, M. J.; Paterson, R. A.; Farnsworth, K. D. and Richardson, D. M. (2013) Advancing impact prediction and hypothesis testing in invasion ecology using a comparative functional response approach. *Biological Invasions* **16** (4), 735-753, doi: 10.1007/s10530-013-0550-8.

Diekmann, J. and Featherman, W. (1998) Assessing Cost Uncertainty: Lessons from Environmental Restoration Projects. *Journal of Construction Engineering and Management* **124** (6), 445-451, doi: 10.1061/(ASCE)0733-9364(1998)124:6(445).

Dierberg, F. E.; DeBusk, T. A.; Jackson, S. D.; Chimney, M. J. and Pietro, K. (2002) Submerged aquatic vegetation-based treatment wetlands for removing phosphorus from agricultural runoff: response to hydraulic and nutrient loading. *Water Research* **36** (6), 1409-1422, doi: 10.1016/S0043-1354(01)00354-2.

Dodd, J. A.; Dick, J. T. A.; Alexander, M. E.; MacNeil, C.; Dunn, A. M. and Aldridge, D. C. (2014) Predicting the ecological impacts of a new freshwater invader: functional responses and prey selectivity of the 'killer shrimp', *Dikerogammarus villosus*, compared to the native *Gammarus pulex*. *Freshwater Biology* **59** (2), 337-352, doi: 10.1111/fwb.12268.

Dodds, W. K. and Whiles, M. R. (2010) *Freshwater Ecology: Concepts and Environmental Applications of Limnology*. 2nd edn (Elsevier. Eastborne, UK).

Döll, P.; Müller Schmied, H.; Schuh, C.; Portmann, F. T. and Eicker, A. (2014) Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research* **50** (7), 5698-5720, doi: 10.1002/2014WR015595.

Domisch, S.; Araújo, M. B.; Bonada, N.; Pauls, S. U.; Jähnig, S. C. and Haase, P. (2013) Modelling distribution in European stream macroinvertebrates under future climates. *Global Change Biology* **19** (3), 752-762, doi: 10.1111/gcb.12107.

Donders, A. R. T.; van der Heijden, G. J. M. G.; Stijnen, T. and Moons, K. G. M. (2006) Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* **59** (10), 1087-1091, doi: 10.1016/j.jclinepi.2006.01.014.

Donoso, N.; Gobeyn, S.; Boets, P.; Goethals, P. L. M.; Wilde, D. D. and Meers, E. (2017) Assessing the Integration of Wetlands along Small European Waterways to Address Diffuse Nitrate Pollution. *Water* **9** (6), doi: 10.3390/w9060369.

Donoso, N.; Gobeyn, S.; Villa-Cox, G.; Boets, P.; Meers, E. and Goethals, P. (2018) Assessing the Ecological Relevance of Organic Discharge Limits for Constructed Wetlands by Means of a Model-Based Analysis. *Water* **10** (1), doi: 10.3390/w10010063.

Donoso, N.; van Oirschot, D.; Kumar Biswas, J.; Michels, E. and Meers, E. (2019) Impact of Aeration on the Removal of Organic Matter and Nitrogen Compounds in Constructed Wetlands Treating the Liquid Fraction of Piggery Manure. *Applied Sciences* **9** (20), doi: 10.3390/app9204310.

Dormann, C. F.; Schymanski, S. J.; Cabral, J.; Chuine, I.; Graham, C.; Hartig, F.; Kearney, M.; Morin, X.; Römermann, C.; Schröder, B. and Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography* **39** (12), 2119-2131, doi: 10.1111/j.1365-2699.2011.02659.x.

Driever, S. M.; Nes, E. H. v. and Roijackers, R. M. M. (2005) Growth limitation of *Lemna minor* due to high plant density. *Aquatic Botany* **81** (3), 245-251, doi: 10.1016/j.aquabot.2004.12.002.

Dubuis, A.; Pottier, J.; Rion, V.; Pellissier, L.; Theurillat, J.-P. and Guisan, A. (2011) Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions* **17** (6), 1122-1131, doi: 10.1111/j.1472-4642.2011.00792.x.

Early, R.; Bradley, B. A.; Dukes, J. S.; Lawler, J. J.; Olden, J. D.; Blumenthal, D. M.; Gonzalez, P.; Grosholz, E. D.; Ibañez, I.; Miller, L. P.; Sorte, C. J. B. and Tatem, A. J. (2016) Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature Communications* **7** (1), 12485, doi: 10.1038/ncomms12485.

Eivers, R. S.; Duggan, I. C.; Hamilton, D. P. and Quinn, J. M. (2017) Constructed treatment wetlands provide habitat for zooplankton communities in agricultural peat lake catchments. *Wetlands*, doi: 10.1007/s13157-017-0959-4.

Elith, J.; Ferrier, S.; Huettmann, F. and Leathwick, J. (2005) The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling* **186** (3), 280-289, doi: 10.1016/j.ecolmodel.2004.12.007.

Elith, J. and Graham, C. H. (2009) Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32** (1), 66-77, doi: 10.1111/j.1600-0587.2008.05505.x.

Elith, J.; Graham, C. H.; Anderson, R. P.; Dudík, M.; Ferrier, S.; Guisan, A.; Hijmans, R. J.; Huettmann, F.; Leathwick, J. R.; Lehmann, A.; Li, J.; Lohmann, L. G.; Loiselle, B. A.; Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; McC. M. Overton, J.; Townsend Peterson, A.; Phillips, S. J.; Richardson, K.; Scachetti-Pereira, R.; Schapire, R. E.; Soberón, J.; Williams, S.; Wisz, M. S. and Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29** (2), 129-151, doi: 10.1111/j.2006.0906-7590.04596.x.

Elith, J. and Leathwick, J. R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics* **40** (1), 677-697, doi: 10.1146/annurev.ecolsys.110308.120159.

Eller, F.; Alnoee, A. B.; Boderskov, T.; Guo, W.-Y.; Kamp, A. T.; Sorrell, B. K. and Brix, H. (2015) Invasive submerged freshwater macrophytes are more plastic in their response to light intensity than to the availability of free CO2 in air-equilibrated water. *Freshwater Biology* **60** (5), 929-943, doi: 10.1111/fwb.12547.

Engelhardt, K. A. M. and Ritchie, M. E. (2001) Effects of macrophyte species richness on wetland ecosystem functioning and services. *Nature* **411** (6838), 687-689, doi: 10.1038/35079573.

Evans, J. S. and Cushman, S. A. (2009) Gradient modeling of conifer species using random forests. *Landscape Ecology* **24** (5), 673-683, doi: 10.1007/s10980-009-9341-0.

Everaert, G.; Boets, P.; Lock, K.; Džeroski, S. and Goethals, P. L. M. (2011) Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium. *Ecological Modelling* **222** (14), 2202-2212, doi: 10.1016/j.ecolmodel.2010.08.013.

Everaert, G.; De Neve, J.; Boets, P.; Dominguez-Granda, L.; Mereta, S. T.; Ambelu, A.; Hoang, T. H.; Goethals, P. L. and Thas, O. (2014) Comparison of the abiotic preferences of macroinvertebrates in tropical river basins. *PLoS ONE* **9** (10), e108898, doi: 10.1371/journal.pone.0108898.

Everaert, G.; Pauwels, I.; Bennetsen, E. and Goethals, P. L. M. (2016) Development and selection of decision trees for water management: Impact of data preprocessing, algorithms and settings. *AI Communications* **29** (6), 711-723, doi: 10.3233/AIC-160711.

Fagúndez, J. and Lema, M. (2019) A competition experiment of an invasive alien grass and two native species: are functionally similar species better competitors? *Biological Invasions* **21** (12), 3619-3631, doi: 10.1007/s10530-019-02073-y.

Fairchild, G. W.; Faulds, A. M. and Matta, J. F. (2000) Beetle assemblages in ponds: effects of habitat and site age. *Freshwater Biology* **44** (3), 523-534, doi: 10.1046/j.1365-2427.2000.00601.x.

Fan, J.; Zhang, J.; Ngo, H. H.; Guo, W. and Yin, X. (2016) Improving low-temperature performance of surface flow constructed wetlands using Potamogeton crispus L. plant. *Bioresource Technology* **218**, 1257-1260, doi: 10.1016/j.biortech.2016.06.110.

Faris, P. D.; Ghali, W. A.; Brant, R.; Norris, C. M.; Galbraith, P. D. and Knudtson, M. L. (2002) Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology* **55** (2), 184-191, doi: 10.1016/S0895-4356(01)00433-4.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters* **27** (8), 861-874, doi: 10.1016/j.patrec.2005.10.010.

Fielding, A. H. and Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24** (01), 38-49, doi: 10.1017/S0376892997000088.

Fontanarrosa, M. S.; Chaparro, G.; de Tezanos Pinto, P.; Rodriguez, P. and O'Farrell, I. (2010) Zooplankton response to shading effects of free-floating plants in shallow warm temperate lakes: a field mesocosm experiment. *Hydrobiologia* **646** (1), 231-242, doi: 10.1007/s10750-010-0183-1.

Forbes, V. E.; Calow, P. and Sibly, R. M. (2008) The extrapolation problem and how population modeling can help. *Environmental Toxicology and Chemistry* **27** (10), 1987-1994, doi: 10.1897/08-029.1.

Forio, M. A. E.; Goethals, P. L. M.; Lock, K.; Asio, V.; Bande, M. and Thas, O. (2018) Model-based analysis of the relationship between macroinvertebrate traits and environmental river conditions. *Environmental Modelling & Software* **106**, 57-67, doi: 10.1016/j.envsoft.2017.11.025.

Forio, M. A. E.; Landuyt, D.; Bennetsen, E.; Lock, K.; Nguyen, T. H. T.; Ambarita, M. N. D.; Musonge, P. L. S.; Boets, P.; Everaert, G.; Dominguez-Granda, L. and Goethals, P. L. M. (2015) Bayesian belief network models to analyse and predict ecological water quality in rivers. *Ecological Modelling* **312**, 222-238, doi: 10.1016/j.ecolmodel.2015.05.025.

Fox, E. W.; Hill, R. A.; Leibowitz, S. G.; Olsen, A. R.; Thornbrugh, D. J. and Weber, M. H. (2017) Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment* **189** (7), 316, doi: 10.1007/s10661-017-6025-0.

Franklin, J. (2010) *Mapping species distributions: spatial inference and prediction*. (Cambridge University Press.).

Frédéric, M.; Samir, L.; Louise, M. and Abdelkrim, A. (2006) Comprehensive modeling of mat density effect on duckweed (Lemna minor) growth under controlled eutrophication. *Water Research* **40** (15), 2901-2910, doi: 10.1016/j.watres.2006.05.026.

Freeman, E. and Moisen, G. (2008a) PresenceAbsence: An R Package for Presence-Absence Model Analysis. *Journal of Statistical Software* **23** (11), 1-31.

Freeman, E. A. and Moisen, G. G. (2008b) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling* **217** (1), 48-58, doi: 10.1016/j.ecolmodel.2008.05.015.

Freeman, E. A.; Moisen, G. G.; Coulston, J. W. and Wilson, B. T. (2015) Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research* **46** (3), 323-339, doi: 10.1139/cjfr-2014-0562.

Friberg, N.; Bonada, N.; Bradley, D. C.; Dunbar, M. J.; Edwards, F. K.; Grey, J.; Hayes, R. B.; Hildrew, A. G.; Lamouroux, N. and Trimmer, M. (2011) Biomonitoring of human impacts in freshwater ecosystems: the good, the bad and the ugly. *Advances in Ecological Research* **44**, 1-68.

Friberg, N.; Buijse, T.; Carter, C.; Hering, D.; M. Spears, B.; Verdonschot, P. and Moe, T. F. (2017) Effective restoration of aquatic ecosystems: scaling the barriers. *WIREs Water* **4** (1), e1190, doi: 10.1002/wat2.1190.

Friedman, N.; Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning* **29** (2-3), 131-163, doi: 10.1023/A:1007465528199.

Frodge, J. D.; Thomas, G. L. and Pauley, G. B. (1990) Effects of canopy formation by floating and submergent aquatic macrophytes on the water quality of two shallow Pacific Northwest lakes. *Aquatic Botany* **38** (2), 231-248, doi: 10.1016/0304-3770(90)90008-9.

Fukuda, S.; De Baets, B.; Mouton, A. M.; Waegeman, W.; Nakajima, J.; Mukai, T.; Hiramatsu, K. and Onikura, N. (2011) Effect of model formulation on the optimization of a genetic Takagi–Sugeno fuzzy system for fish habitat suitability evaluation. *Ecological Modelling* **222** (8), 1401-1413, doi: 10.1016/j.ecolmodel.2011.01.023.

Fukuda, S.; De Baets, B.; Waegeman, W.; Verwaeren, J. and Mouton, A. M. (2013) Habitat prediction and knowledge extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species distribution models. *Environmental Modelling & Software* **47**, 1-6, doi: 10.1016/j.envsoft.2013.04.005.

Funk, J. L. and Vitousek, P. M. (2007) Resource-use efficiency and plant invasion in low-resource systems. *Nature* **446** (7139), 1079-1081, doi: 10.1038/nature05719.

Galanopoulos, C.; Sazakli, E.; Leotsinidis, M. and Lyberatos, G. (2013) A pilot-scale study for modeling a free water surface constructed wetlands wastewater treatment system. *Journal of Environmental Chemical Engineering* **1** (4), 642-651, doi: 10.1016/j.jece.2013.09.006.

Galatowitsch, S. M. (2006) Restoring prairie pothole wetlands: does the species pool concept offer decision-making guidance for re-vegetation? *Applied Vegetation Science* **9** (2), 261-270, doi: 10.1658/1402-2001(2006)9[261:RPPWDT]2.0.CO;2.

Gallardo, B. and Aldridge, D. C. (2013) Evaluating the combined threat of climate change and biological invasions on endangered species. *Biological Conservation* **160** (0), 225-233, doi: 10.1016/j.biocon.2013.02.001.

Gallardo, B.; Errea, M. P. and Aldridge, D. (2012) Application of bioclimatic models coupled with network analysis for risk assessment of the killer shrimp, Dikerogammarus villosus, in Great Britain. *Biological Invasions* **14** (6), 1265-1278, doi: 10.1007/s10530-011-0154-0.

Gallien, L.; Douzet, R.; Pratte, S.; Zimmermann, N. E. and Thuiller, W. (2012) Invasive species distribution models – how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography* **21** (11), 1126-1136, doi: 10.1111/j.1466-8238.2012.00768.x.

Gallien, L.; Münkemüller, T.; Albert, C. H.; Boulangeat, I. and Thuiller, W. (2010) Predicting potential distributions of invasive species: where to go from here? *Diversity and Distributions* **16** (3), 331-342, doi: 10.1111/j.1472-4642.2010.00652.x.

Gao, X.; Wang, Y.; Sun, B. and Li, N. (2019) Nitrogen and phosphorus removal comparison between periphyton on artificial substrates and plant-periphyton complex in floating treatment wetlands. *Environmental Science and Pollution Research* **26** (21), 21161-21171, doi: 10.1007/s11356-019-05455-w.

Gao, Y.; Xie, Y. W.; Zhang, Q.; Wang, A. L.; Yu, Y. X. and Yang, L. Y. (2017) Intensified nitrate and phosphorus removal in an electrolysis -integrated horizontal subsurface-flow constructed wetland. *Water Research* **108**, 39-45, doi: 10.1016/j.watres.2016.10.033.

García-Laencina, P. J.; Sancho-Gómez, J.-L. and Figueiras-Vidal, A. R. (2010) Pattern classification with missing data: a review. *Neural Computing and Applications* **19** (2), 263-282, doi: 10.1007/s00521-009-0295-6.

García-Lledó, A.; Ruiz-Rueda, O.; Vilar-Sanz, A.; Sala, L. and Bañeras, L. (2011) Nitrogen removal efficiencies in a free water surface constructed wetland in relation to plant coverage. *Ecological Engineering* **37** (5), 678-684, doi: 10.1016/j.ecoleng.2010.06.034.

Garfí, M.; Pedescoll, A.; Bécares, E.; Hijosa-Valsero, M.; Sidrach-Cardona, R. and García, J. (2012) Effect of climatic conditions, season and wastewater quality on contaminant removal efficiency of two experimental constructed wetlands in different regions of Spain. *Science of the Total Environment* **437**, 61-67, doi: 10.1016/j.scitotenv.2012.07.087.

Genkai-Kato, M. and Carpenter, S. R. (2005) Eutrophication due to phosphorus recycling in relation to lake morphometry, temperature, and macrophytes. *Ecology* **86** (1), 210-219, doi: 10.1890/03-0545.

Gérard, J. and Triest, L. (2014) The Effect of Phosphorus Reduction and Competition on Invasive Lemnids: Life Traits and Nutrient Uptake. *ISRN Botany* **2014**, 9, doi: 10.1155/2014/514294.

Giannini, T. C.; Chapman, D. S.; Saraiva, A. M.; Alves-dos-Santos, I. and Biesmeijer, J. C. (2013) Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography* **36** (6), 649-656, doi: 10.1111/j.1600-0587.2012.07191.x.

Gibert, K.; Horsburgh, J. S.; Athanasiadis, I. N. and Holmes, G. (2018a) Environmental Data Science. *Environmental Modelling & Software* **106**, 4-12, doi: 10.1016/j.envsoft.2018.04.005.

Gibert, K.; Izquierdo, J.; Sànchez-Marrè, M.; Hamilton, S. H.; Rodríguez-Roda, I. and Holmes, G. (2018b) Which method to use? An assessment of data mining methods in Environmental Data Science. *Environmental Modelling & Software* **110**, 3-27, doi: 10.1016/j.envsoft.2018.09.021.

Gioria, M. and Osborne, B. A. (2014) Resource competition in plant invasions: emerging patterns and research needs. *Frontiers in Plant Science* **5**, 501, doi: 10.3389/fpls.2014.00501.

Giustarini, L.; Parisot, O.; Ghoniem, M.; Hostache, R.; Trebs, I. and Otjacques, B. (2016) A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. *Environmental Modelling & Software* **82**, 308-320, doi: 10.1016/j.envsoft.2016.04.013.

Gobeyn, S.; Volk, M.; Dominguez-Granda, L. and Goethals, P. L. M. (2017) Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador. *Environmental Modelling & Software* **92**, 269-316, doi: 10.1016/j.envsoft.2017.02.012.

Goethals, P. L. M. *Data driven development of predictive ecological models for benthic macroinvertebrates in rivers* PhD thesis, Ghent, (2005).

Goethals, P. L. M.; Dedecker, A. P.; Gabriels, W.; Lek, S. and De Pauw, N. (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology* **41** (3), 491-508, doi: 10.1007/s10452-007-9093-3.

Gooden, B. and French, K. (2015) Impacts of alien plant invasion on native plant communities are mediated by functional identity of resident species, not resource availability. *Oikos* **124** (3), 298-306, doi: 10.1111/oik.01724.

Gopal, B. (2016) Should 'wetlands' cover all aquatic ecosystems and do macrophytes make a difference to their ecosystem services? *Folia Geobotanica* **51** (3), 209-226, doi: 10.1007/s12224-016-9248-x.

Gregorutti, B.; Michel, B. and Saint-Pierre, P. (2017) Correlation and variable importance in random forests. *Statistics and Computing* **27** (3), 659-678, doi: 10.1007/s11222-016-9646-1.

Grotkopp, E.; Rejmánek, M. and Rost, T. L. (2002) Toward a Causal Explanation of Plant Invasiveness: Seedling Growth and Life-History Strategies of 29 Pine (*Pinus*) Species. *The American Naturalist* **159** (4), 396-419, doi: 10.1086/338995.

Gu, W. and Swihart, R. K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation* **116** (2), 195-203, doi: 10.1016/S0006-3207(03)00190-3.

Gueta, T. and Carmel, Y. (2016) Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics* **34**, 139-145, doi: 10.1016/j.ecoinf.2016.06.001.

Guillaume, S. (2001) Designing fuzzy inference systems from data: An interpretability-oriented review. *Fuzzy Systems, IEEE Transactions on* **9** (3), 426-443, doi: 10.1109/91.928739.

Guisan, A.; Edwards Jr, T. C. and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157** (2–3), 89-100, doi: 10.1016/S0304-3800(02)00204-1.

Guisan, A.; Lehmann, A.; Ferrier, S.; Austin, M.; Overton, J. M. C.; Aspinall, R. and Hastie, T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology* **43** (3), 386-392, doi: 10.1111/j.1365-2664.2006.01164.x.

Guisan, A. and Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography* **38** (8), 1433-1444, doi: 10.1111/j.1365-2699.2011.02550.x.

Guisan, A. and Theurillat, J.-P. (2000) Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia* **30** (3/4), 353-384.

Guisan, A. and Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters* **8** (9), 993-1009, doi: 10.1111/j.1461-0248.2005.00792.x.

Guisan, A. and Zimmerman, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* **135**, 147-186, doi: 10.1016/S0304-3800(00)00354-9.

Guo, C.; Lek, S.; Ye, S.; Li, W.; Liu, J. and Li, Z. (2015) Uncertainty in ensemble modelling of large-scale species distribution: Effects from species characteristics and model techniques. *Ecological Modelling* **306**, 67-75, doi: 10.1016/j.ecolmodel.2014.08.002.

Haddaway, N. R.; Wilcox, R. H.; Heptonstall, R. E. A.; Griffiths, H. M.; Mortimer, R. J. G.; Christmas, M. and Dunn, A. M. (2012) Predatory Functional Response and Prey Choice Identify Predation Differences between Native/Invasive and Parasitised/Unparasitised Crayfish. *PLoS ONE* **7** (2), e32229, doi: 10.1371/journal.pone.0032229.

Hammouda, O.; Gaber, A. and Abdel-Hameed, M. S. (1995) Assessment of the effectiveness of treatment of wastewater-contaminated aquatic systems with *Lemna gibba*. *Enzyme and Microbial Technology* **17** (4), 317-323, doi: 10.1016/0141-0229(94)00013-1.

Hansson, L.-A.; Brönmark, C.; Anders Nilsson, P. and Åbjörnsson, K. (2005) Conflicting demands on wetland ecosystem services: nutrient retention, biodiversity or both? *Freshwater Biology* **50** (4), 705-714, doi: 10.1111/j.1365-2427.2005.01352.x.

Hardin, G. (1968) The Tragedy of the Commons. *Science* **162** (3859), 1243, doi: 10.1126/science.162.3859.1243.

Harrel, F. E. J. (2018) *Hmisc: Harrel Miscellaneous* v. 4.1-1.

Harrington, R. and McInnes, R. (2009) Integrated Constructed Wetlands (ICW) for livestock wastewater management. *Bioresource Technology* **100** (22), 5498-5505, doi: 10.1016/j.biortech.2009.06.007.

Harris, J. A.; Hobbs, R. J.; Higgs, E. and Aronson, J. (2006) Ecological Restoration and Global Climate Change. *Restoration Ecology* **14** (2), 170-176, doi: 10.1111/j.1526-100X.2006.00136.x.

Harrison, I.; Abell, R.; Darwall, W.; Thieme, M. L.; Tickner, D. and Timboe, I. (2018) The freshwater biodiversity crisis. *Science* **362** (6421), 1369, doi: 10.1126/science.aav9242.

Hartig, T.; Mitchell, R.; de Vries, S. and Frumkin, H. (2014) Nature and Health. *Annual Review of Public Health* **35** (1), 207-228, doi: 10.1146/annurev-publhealth-032013-182443.

Hastie, T.; Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd Edition edn (Springer. Stanford, California).

Hatten, J.; Batt, T.; Connolly, P. and Maule, A. (2014) Modeling effects of climate change on Yakima River salmonid habitats. *Climatic Change* **124** (1-2), 427-439, doi: 10.1007/s10584-013-0980-4.

He, F.; Zarfl, C.; Bremerich, V.; David, J. N. W.; Hogan, Z.; Kalinkat, G.; Tockner, K. and Jähnig, S. C. (2019) The global decline of freshwater megafauna. *Global Change Biology* **25** (11), 3883-3892, doi: 10.1111/gcb.14753.

He, H.; Chen, Y.; Li, X.; Cheng, Y.; Yang, C. and Zeng, G. (2017) Influence of salinity on microorganisms in activated sludge processes: A review. *International Biodeterioration & Biodegradation* **119**, 520-527, doi: 10.1016/j.ibiod.2016.10.007.

Healy, M. G.; Rodgers, M. and Mulqueen, J. (2007) Treatment of dairy wastewater using constructed wetlands and intermittent sand filters. *Bioresource Technology* **98** (12), 2268-2281, doi: 10.1016/j.biortech.2006.07.036.

Hefting, M. M.; van den Heuvel, R. N. and Verhoeven, J. T. A. (2013) Wetlands in agricultural landscapes for nitrogen attenuation and biodiversity enhancement: Opportunities and limitations. *Ecological Engineering* **56**, 5-13, doi: 10.1016/j.ecoleng.2012.05.001.

Henry-Silva, G. G.; Camargo, A. F. M. and Pezzato, M. M. (2008) Growth of free-floating aquatic macrophytes in different concentrations of nutrients. *Hydrobiologia* **610** (1), 153-160, doi: 10.1007/s10750-008-9430-0.

Herbert, E. R.; Boon, P.; Burgin, A. J.; Neubauer, S. C.; Franklin, R. B.; Ardón, M.; Hopfensperger, K. N.; Lamers, L. P. M. and Gell, P. (2015) A global perspective on wetland salinization: ecological consequences of a growing threat to freshwater wetlands. *Ecosphere* **6** (10), 1-43, doi: 10.1890/ES14-00534.1.

Hernández-Crespo, C.; Gargallo, S.; Benedito-Durá, V.; Nácher-Rodríguez, B.; Rodrigo-Alacreu, M. A. and Martín, M. (2017) Performance of surface and subsurface flow constructed wetlands treating eutrophic waters. *Science of the Total Environment* **595** (Supplement C), 584-593, doi: 10.1016/j.scitotenv.2017.03.278.

Hernández-Crespo, C.; Oliver, N.; Bixquert, J.; Gargallo, S. and Martín, M. (2016) Comparison of three plants in a surface flow constructed wetland treating eutrophic water in a Mediterranean climate. *Hydrobiologia* **774** (1), 183-192, doi: 10.1007/s10750-015-2493-9.

Hernández, M. A. and Stolfo, S. J. (1998) Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery* **2** (1), 9-37, doi: 10.1023/A:1009761603038.

Hijosa-Valsero, M.; Sidrach-Cardona, R.; Martín-Villacorta, J. and Bécares, E. (2010) Optimization of performance assessment and design characteristics in constructed wetlands for the removal of organic matter. *Chemosphere* **81** (5), 651-657, doi: 10.1016/j.chemosphere.2010.08.010.

Hilderbrand, R.; Watts, A. and Randle, A. (2005) The myths of restoration ecology. *Ecology and Society* **10** (1).

Hillman, W. S. (1961) The Lemnaceae, or duckweeds. *The Botanical Review* **27** (2), 221-287, doi: 10.1007/BF02860083.

Hilt, S.; Gross, E. M.; Hupfer, M.; Morscheid, H.; Mählmann, J.; Melzer, A.; Poltz, J.; Sandrock, S.; Scharf, E.-M.; Schneider, S. and van de Weyer, K. (2006) Restoration of submerged vegetation in shallow eutrophic lakes – A guideline and state of the art in Germany. *Limnologica - Ecology and Management of Inland Waters* **36** (3), 155-171, doi: 10.1016/j.limno.2006.06.001.

Hirzel, A. H.; Hausser, J.; Chessel, D. and Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* **83** (7), 2027-2036, doi: 10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2.

Hoang, T. H.; Lock, K.; Mouton, A. and Goethals, P. L. M. (2010) Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecological Informatics* **5** (2), 140-146, doi: 10.1016/j.ecoinf.2009.12.001.

Hofstra, D.; Schoelynck, J.; Ferrell, J.; Coetzee, J.; de Winton, M.; Bickel, T. O.; Champion, P.; Madsen, J.; Bakker, E. S.; Hilt, S.; Matheson, F.; Netherland, M. and Gross, E. M. (2020) On the move: New insights on the ecology and management of native and alien macrophytes. *Aquatic Botany* **162**, 103190, doi: 10.1016/j.aquabot.2019.103190.

Hothorn, T.; Bretz, F. and Westfall, P. (2008) Simultaneous Inference in General Parametric Models. *Biometrical Journal* **50** (3), 346-363, doi: 10.1002/bimj.200810425.

Hothorn, T.; Hornik, K.; Strobl, C. and Zeileis, A. (2018) *party: A Laboratory for Recursive Partitioning* v. 1.3-1.

Hsu, C.-B.; Hsieh, H.-L.; Yang, L.; Wu, S.-H.; Chang, J.-S.; Hsiao, S.-C.; Su, H.-C.; Yeh, C.-H.; Ho, Y.-S. and Lin, H.-J. (2011) Biodiversity of constructed wetlands for wastewater treatment. *Ecological Engineering* **37** (10), 1533-1545, doi: 10.1016/j.ecoleng.2011.06.002.

Huntley, B.; Green, R. E.; Collingham, Y. C.; Hill, J. K.; Willis, S. G.; Bartlein, P. J.; Cramer, W.; Hagemeijer, W. J. M. and Thomas, C. J. (2004) The performance of models relating species geographical distributions to climate is independent of trophic level. *Ecology Letters* **7** (5), 417-426, doi: 10.1111/j.1461-0248.2004.00598.x.

Hussner, A. (2009) Growth and photosynthesis of four invasive aquatic plant species in Europe. *Weed Research* **49** (5), 506-515, doi: 10.1111/j.1365-3180.2009.00721.x.

Hussner, A. (2012) Alien aquatic plant species in European countries. *Weed Research* **52** (4), 297-306, doi: 10.1111/j.1365-3180.2012.00926.x.

Hutchinson, G. E. (1957) Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology* **22**, 415-427, doi: 10.1101/SQB.1957.022.01.039.

IPCC. (2014) *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (IPCC, Geneva, Switzerland).

Ishida, C. K.; Arnon, S.; Peterson, C. G.; Kelly, J. J. and Gray, K. A. (2008) Influence of Algal Community Structure on Denitrification Rates in Periphyton Cultivated on Artificial Substrata. *Microbial Ecology* **56** (1), 140-152, doi: 10.1007/s00248-007-9332-0.

IUCN. (2019) *Global Invasive Species Database*, <www.iucngisd.org/gisd/> (Last accessed on 29/10/2019).

Iverson, L. R. and Prasad, A. M. (1998) Predicting abundance of 80 tree species following climate change in the Eastern United States. *Ecological Monographs* **68** (4), 465-485, doi: 10.1890/0012-9615(1998)068[0465:PAOTSF]2.0.CO;2.

Jackson, L. L.; Lopoukhine, N. and Hillyard, D. (1995) Ecological Restoration: A Definition and Comments. *Restoration Ecology* **3** (2), 71-75, doi: 10.1111/j.1526-100X.1995.tb00079.x.

Jackson, S. T. and Hobbs, R. J. (2009) Ecological Restoration in the Light of Ecological History. *Science* **325** (5940), 567, doi: 10.1126/science.1172977.

Jähnig, S. C.; Lorenz, A. W.; Hering, D.; Antons, C.; Sundermann, A.; Jedicke, E. and Haase, P. (2011) River restoration success: a question of perception. *Ecological Applications* **21** (6), 2007-2015, doi: 10.1890/10-0618.1.

Jakeman, A. J.; Letcher, R. A. and Norton, J. P. (2006) Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software* **21** (5), 602-614, doi: 10.1016/j.envsoft.2006.01.004.

Janes, R. A.; Eaton, J. W. and Hardwick, K. (1996) The effects of floating mats of *Azolla filiculoides* Lam. and *Lemna minuta* Kunth on the growth of submerged macrophytes. *Hydrobiologia* **340** (1), 23-26, doi: 10.1007/BF00012729.

Janse, J. H. and Van Puijenbroek, P. J. T. M. (1998) Effects of eutrophication in drainage ditches. *Environmental Pollution* **102** (1, Supplement 1), 547-552, doi: 10.1016/S0269-7491(98)80082-1.

Jarchow, M. E. and Cook, B. J. (2009) Allelopathy as a mechanism for the invasion of Typha angustifolia. *Plant Ecology* **204** (1), 113-124, doi: 10.1007/s11258-009-9573-8.

Jarnevich, C. S.; Stohlgren, T. J.; Kumar, S.; Morisette, J. T. and Holcombe, T. R. (2015) Caveats for correlative species distribution modeling. *Ecological Informatics* **29**, 6-15, doi: 10.1016/j.ecoinf.2015.06.007.

Jenkins, W. A.; Murray, B. C.; Kramer, R. A. and Faulkner, S. P. (2010) Valuing ecosystem services from wetlands restoration in the Mississippi Alluvial Valley. *Ecological Economics* **69** (5), 1051-1061, doi: 10.1016/j.ecolecon.2009.11.022.

Jensen, F. V. and Nielsen, T. D. (2007) *Bayesian network and decision graphs*. 2nd edn, Vol. XVI (Springer Verlag.).

Jiménez-Valverde, A. and Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions* **12** (5), 521-524, doi: 10.1111/j.1366-9516.2006.00267.x.

Jiménez-Valverde, A.; Lobo, J. M. and Hortal, J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions* **14** (6), 885-890, doi: 10.1111/j.1472-4642.2008.00496.x.

Junk, W. J.; Piedade, M. T. F.; Lourival, R.; Wittmann, F.; Kandus, P.; Lacerda, L. D.; Bozelli, R. L.; Esteves, F. A.; Nunes da Cunha, C.; Maltchik, L.; Schöngart, J.; Schaeffer-Novelli, Y. and Agostinho, A. A. (2014) Brazilian wetlands: their definition, delineation, and classification for research, sustainable management, and protection. *Aquatic Conservation: Marine and Freshwater Ecosystems* **24** (1), 5-22, doi: 10.1002/aqc.2386.

Kadlec, R. H. (2009) Comparison of free water and horizontal subsurface treatment wetlands. *Ecological Engineering* **35** (2), 159-174, doi: 10.1016/j.ecoleng.2008.04.008.

Kadlec, R. H. and Wallace, S. (2008) *Treatment wetlands*. 2nd edn (CRC press. Boca Raton, Florida, USA).

Kampf, R. and Claassen, T. H. L. (2004) The use of treated wastewater for nature: the Waterharmonica, a sustainable solution as an alternative for separate drainage and treatment in *2nd IWA Leading-Edge on Water and Wastewater Treatment Technologies Water and Environmental Management* (eds M. Van Loosdrecht and J. Clement) 341 (IWA Publishing, London, UK).

Kampichler, C.; Wieland, R.; Calmé, S.; Weissenberger, H. and Arriaga-Weiss, S. (2010) Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics* **5** (6), 441-450, doi: 10.1016/j.ecoinf.2010.06.003.

Karathanasis, A. D.; Potter, C. L. and Coyne, M. S. (2003) Vegetation effects on fecal bacteria, BOD, and suspended solid removal in constructed wetlands treating domestic wastewater. *Ecological Engineering* **20** (2), 157-169, doi: 10.1016/s0925-8574(03)00011-9.

Kassambra, A. (2019) *ggpubr: 'ggplot2' Based Publication Ready Plots* v. R package version 0.2.2.

Kearney, M. and Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters* **12** (4), 334-350, doi: 10.1111/j.1461-0248.2008.01277.x.

Keddy, P. (1999) Wetland restoration: The potential for assembly rules in the service of conservation. *Wetlands* **19** (4), 716-732, doi: 10.1007/BF03161780.

Keddy, P. A. (2010) *Wetland ecology: principles and conservation*. 2nd edn (Cambridge University Press. New York, USA).

Kelly, R.; Harrod, C.; Maggs, C. A. and Reid, N. (2015) Effects of Elodea nuttallii on temperate freshwater plants, microalgae and invertebrates: small differences between invaded and uninvaded areas. *Biological Invasions* **17** (7), 2123-2138, doi: 10.1007/s10530-015-0865-8.

Keshtkar, A. R.; Salajegheh, A.; Sadoddin, A. and Allan, M. G. (2013) Application of Bayesian networks for sustainability assessment in catchment modeling and management (Case study: The Hablehrood river catchment). *Ecological Modelling* **268** (0), 48-54, doi: 10.1016/j.ecolmodel.2013.08.003.

Kingsford, R. T.; Basset, A. and Jackson, L. (2016) Wetlands: conservation's poor cousins. *Aquatic Conservation: Marine and Freshwater Ecosystems* **26** (5), 892-916, doi: 10.1002/aqc.2709.

Kivaisi, A. K. (2001) The potential for constructed wetlands for wastewater treatment and reuse in developing countries: a review. *Ecological Engineering* **16** (4), 545-560, doi: 10.1016/S0925-8574(00)00113-0.

Kjærulff, U. (1995) dHugin: a computational system for dynamic time-sliced Bayesian networks. *International Journal of Forecasting* **11** (1), 89-111, doi: 10.1016/0169-2070(94)02003-8.

Klir, G. J. and Yuan, B. O. (1995) *Fuzzy sets and fuzzy logic*. (Prentice Hall PTR. Upper Sadle River, New Jersey, 07458, USA).

Knight, R. L.; Clarke, R. A. and Bastian, R. K. (2001) Surface flow (SF) treatment wetlands as a habitat for wildlife and humans. *Water Science and Technology* **44** (11-12), 27, doi: 10.2166/wst.2001.0806.

Knoben, R. and van der Wal, B. in *Occurrence Dataset* (ed D. F. f. A. W. Research) (2015).

Kolar, C. S. and Lodge, D. M. (2002) Ecological Predictions and Risk Assessment for Alien Fishes in North America. *Science* **298** (5596), 1233-1236, doi: 10.1126/science.1075753.

Kompare, B.; Bratko, I.; Steinman, F. and Džeroski, S. (1994) Using machine learning techniques in the construction of models I. Introduction. *Ecological Modelling* **75–76** (0), 617-628, doi: 10.1016/0304-3800(94)90054-X.

Körner, S. and Dugdale, T. (2003) Is roach herbivory preventing re-colonization of submerged macrophytes in a shallow lake? *Hydrobiologia* **506** (1), 497-501, doi: 10.1023/B:HYDR.0000008561.67513.ec.

Körner, S. and Vermaat, J. E. (1998) The relative importance of Lemna gibba L., bacteria and algae for the nitrogen and phosphorus removal in duckweed-covered domestic wastewater. *Water Research* **32** (12), 3651-3661, doi: 10.1016/S0043-1354(98)00166-3.

Kotsiantis, S. B. (2011) Decision trees: a recent overview. *Artificial Intelligence Review* **39** (4), 261-283, doi: 10.1007/s10462-011-9272-4.

Kotsiantis, S. B.; Kanellopoulos, D. and Pintelas, P. E. (2006) Data preprocessing for supervised leaning. *International Journal of Computer Science* **1** (2), 111-117.

Kotti, I. P.; Gikas, G. D. and Tsihrintzis, V. A. (2010) Effect of operational and design parameters on removal efficiency of pilot-scale FWS constructed wetlands and comparison with HSF systems. *Ecological Engineering* **36** (7), 862-875, doi: 10.1016/j.ecoleng.2010.03.002.

Kovalenko, K. E.; Dibble, E. D. and Slade, J. G. (2010) Community effects of invasive macrophyte control: role of invasive plant abundance and habitat complexity. *Journal of Applied Ecology* **47** (2), 318-328, doi: 10.1111/j.1365-2664.2009.01768.x.

Kowarik, A. and Templ, M. (2016) Imputation with the R Package VIM. *Journal of Statistical Software* **74** (7), 1-16, doi: 10.18637/jss.v074.i07.

Kundzewicz, Z. W.; Kanae, S.; Seneviratne, S. I.; Handmer, J.; Nicholls, N.; Peduzzi, P.; Mechler, R.; Bouwer, L. M.; Arnell, N.; Mach, K.; Muir-Wood, R.; Brakenridge, G. R.; Kron, W.; Benito, G.; Honda, Y.; Takahashi, K. and Sherstyukov, B. (2014) Flood risk and climate change: global and regional perspectives. *Hydrological Sciences Journal* **59** (1), 1-28, doi: 10.1080/02626667.2013.857411.

Kuznetsova, A.; Brockhoff, P. B. and Christensen, R. H. B. (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82** (13), 1-26, doi: 10.18637/jss.v082.i13.

Lambinon, J.; De Langhe, J. E.; Delvosalle, L. and Vanhecke, L. (1998) *Flora van België, het Groothertogdom Luxemburg, Noord-Frankrijk en de aangrenzende gebieden : pteridofyten en spermatofyten*. (Meise : Nationale plantentuin van België.).

Landuyt, D.; Broekx, S.; D'Hondt, R.; Engelen, G.; Aertsens, J. and Goethals, P. L. M. (2013) A review of Bayesian belief networks in ecosystem service modelling. *Environmental Modelling & Software* **46** (0), 1-11, doi: 10.1016/j.envsoft.2013.03.011.

Lawson, C. R.; Hodgson, J. A.; Wilson, R. J. and Richards, S. A. (2014) Prevalence, thresholds and the performance of presence–absence models. *Methods in Ecology and Evolution* **5** (1), 54-64, doi: 10.1111/2041-210X.12123.

Leathwick, J. R. (1998) Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science* **9** (5), 719-732.

Lee, K. J. and Carlin, J. B. (2010) Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology* **171** (5), 624-632, doi: 10.1093/aje/kwp425.

Lehmann, A. (1998) GIS modeling of submerged macrophyte distribution using Generalized Additive Models. *Plant Ecology* **139** (1), 113-124, doi: 10.1023/A:1009754417131.

Lek, S. and Guégan, J. F. (1999) Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* **120** (2–3), 65-73, doi: 10.1016/S0304-3800(99)00092-7.

Leohle, C. (1983) Evaluation of theories and calculation tools in ecology. *Ecological Modelling* **19** (4), 239-247, doi: 10.1016/0304-3800(83)90041-8.

Levine, J. M.; Adler, P. B. and Yelenik, S. G. (2004) A meta-analysis of biotic resistance to exotic plant invasions. *Ecology Letters* **7** (10), 975-989, doi: 10.1111/j.1461-0248.2004.00657.x.

Levine, J. M.; Vilà, M.; Antonio, C. M. D.; Dukes, J. S.; Grigulis, K. and Lavorel, S. (2003) Mechanisms underlying the impacts of exotic plant invasions. *Proceedings of the Royal Society of London B: Biological Sciences* **270** (1517), 775-781.

Li, D.; Deogun, J.; Spaulding, W. and Shuart, B. in *Rough Sets and Current Trends in Computing.* (eds S. Tsumoto, R. Słowiński, J. Komorowski and J. W. Grzymała-Busse) 573-579 (Springer Berlin Heidelberg).

Liboriussen, L.; Jeppesen, E.; Bramm, M. E. and Lassen, M. F. (2005) Periphyton-macroinvertebrate interactions in light and fish manipulated enclosures in a clear and a turbid shallow lake. *Aquatic Ecology* **39** (1), 23-39, doi: 10.1007/s10452-004-3039-9.

Liew, A. W.-C.; Law, N.-F. and Yan, H. (2011) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics* **12** (5), 498-513, doi: 10.1093/bib/bbq080.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analyis with missing data*. 2nd edn (John Wiley & Sons, Inc. Hoboken, New Jersey, United States of America).

Liu, J.; Danneels, B.; Vanormelingen, P. and Vyverman, W. (2016) Nutrient removal from horticultural wastewater by benthic filamentous algae Klebsormidium sp., Stigeoclonium spp. and their communities: From laboratory flask to outdoor Algal Turf Scrubber (ATS). *Water Research* **92**, 61-68, doi: 10.1016/j.watres.2016.01.049.

Liu, J.; Wu, Y.; Wu, C.; Muylaert, K.; Vyverman, W.; Yu, H.-Q.; Muñoz, R. and Rittmann, B. (2017) Advanced nutrient removal from surface water by a consortium of attached microalgae and bacteria: A review. *Bioresource Technology* **241**, 1127-1137, doi: 10.1016/j.biortech.2017.06.054.

Liu, S.; Dai, H. and Gan, M. (2018) Information-decomposition-model-based missing value estimation for not missing at random dataset. *International Journal of Machine Learning and Cybernetics* **9** (1), 85-95, doi: 10.1007/s13042-015-0354-5.

Lobo, J. M.; Jiménez-Valverde, A. and Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography* **33** (1), 103-114, doi: 10.1111/j.1600-0587.2009.06039.x.

Lobo, J. M.; Jiménez-Valverde, A. and Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17** (2), 145-151, doi: 10.1111/j.1466-8238.2007.00358.x.

Lorenz, A. W.; Jahnig, S. C. and Hering, D. (2009) Re-meandering German lowland streams: qualitative and quantitative effects of restoration measures on hydromorphology and macroinvertebrates. *Environmental Management* **44** (4), 745-754, doi: 10.1007/s00267-009-9350-4.

Lu, J.; Wang, H.; Pan, M.; Xia, J.; Xing, W. and Liu, G. (2012) Using sediment seed banks and historical vegetation change data to develop restoration criteria for a eutrophic lake in China. *Ecological Engineering* **39**, 95-103, doi: 10.1016/j.ecoleng.2011.11.006.

Ludyanskiy, M. L.; McDonald, D. and MacNeill, D. (1993) Impact of the Zebra Mussel, a Bivalve Invader. *Bioscience* **43** (8), 533-544, doi: 10.2307/1311948.

Luengo, J.; García, S. and Herrera, F. (2012) On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems* **32** (1), 77-108, doi: 10.1007/s10115-011-0424-2.

Lundgren, E. J.; Ramp, D.; Ripple, W. J. and Wallach, A. D. (2018) Introduced megafauna are rewilding the Anthropocene. *Ecography* **41** (6), 857-866, doi: 10.1111/ecog.03430.

Luyiga, S. and Kiwanuka, S. (2003) Plankton composition, distribution and significance in a tropical integrated pilot constructed treatment wetland in Uganda. *Water Science and Technology* **48** (5), 241.

Madley-Dowd, P.; Hughes, R.; Tilling, K. and Heron, J. (2019) The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology* **110**, 63-73, doi: 10.1016/j.jclinepi.2019.02.016.

Maine, M. A.; Suñe, N.; Hadad, H.; Sánchez, G. and Bonetto, C. (2007) Removal efficiency of a constructed wetland for wastewater treatment according to vegetation dominance. *Chemosphere* **68** (6), 1105-1113, doi: 10.1016/j.chemosphere.2007.01.064.

Maldonado, C.; Molina, C. I.; Zizka, A.; Persson, C.; Taylor, C. M.; Albán, J.; Chilquillo, E.; Rønsted, N. and Antonelli, A. (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography* **24** (8), 973-984, doi: 10.1111/geb.12326.

Mamdani, E. H. (1974) Application of fuzzy algorithms for control of simple dynamic plant. *Proceedings of the Institution of Electrical Engineers* **121** (12), 1585-1588.

Manel, S.; Williams, H. C. and Ormerod, S. J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* **38** (5), 921-931, doi: 10.1046/j.1365-2664.2001.00647.x.

Marcot, B. G.; Holthausen, R. S.; Raphael, M. G.; Rowland, M. M. and Wisdom, M. J. (2001) Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management* **153** (1–3), 29-42, doi: 10.1016/S0378-1127(01)00452-2.

Marmion, M.; Parviainen, M.; Luoto, M.; Heikkinen, R. K. and Thuiller, W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* **15** (1), 59-69, doi: 10.1111/j.1472-4642.2008.00491.x.

Marvin, L. B. and John, F. K. (2003) Data mining and the impact of missing data. *Industrial Management & Data Systems* **103** (8), 611-621, doi: 10.1108/02635570310497657.

Matsuzaki, S.-i. S.; Usio, N.; Takamura, N. and Washitani, I. (2008) Contrasting impacts of invasive engineers on freshwater ecosystems: an experiment and meta-analysis. *Oecologia* **158** (4), 673-686, doi: 10.1007/s00442-008-1180-1.

Mayeux, R. (2004) Biomarkers: Potential uses and limitations. *NeuroRX* **1** (2), 182-188, doi: 10.1602/neurorx.1.2.182.

McPherson, J. M.; Jetz, W. and Rogers, D. J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* **41** (5), 811-823, doi: 10.1111/j.0021-8901.2004.00943.x.

Médoc, V.; Albert, H. and Spataro, T. (2015) Functional response comparisons among freshwater amphipods: ratio-dependence and higher predation for *Gammarus pulex* compared to the non-natives *Dikerogammarus villosus* and *Echinogammarus berilloni*. *Biological Invasions* **17** (12), 3625-3637, doi: 10.1007/s10530-015-0984-2.

Meerhoff, M.; Iglesias, C.; De Mello, F. T.; Clemente, J. M.; Jensen, E.; Lauridsen, T. L. and Jeppesen, E. (2007) Effects of habitat complexity on community structure and predator avoidance behaviour of littoral zooplankton in temperate versus subtropical shallow lakes. *Freshwater Biology* **52** (6), 1009-1021, doi: 10.1111/j.1365-2427.2007.01748.x.

Mereta, S. T.; Boets, P.; Ambelu Bayih, A.; Malu, A.; Ephrem, Z.; Sisay, A.; Endale, H.; Yitbarek, M.; Jemal, A.; De Meester, L. and Goethals, P. L. M. (2012) Analysis of environmental factors determining the abundance and diversity of macroinvertebrate taxa in natural wetlands of Southwest Ethiopia. *Ecological Informatics* **7** (1), 52-61, doi: 10.1016/j.ecoinf.2011.11.005.

Merlin, G.; Pajean, J.-L. and Lissolo, T. (2002) Performances of constructed wetlands for municipal wastewater treatment in rural mountainous areas. *Hydrobiologia* **469**, 87-98, doi: 10.1023/A:1015567325463.

Microsoft Corporation and Weston, S. (2019a) *doParallel: Foreach Parallel Adaptor for the 'parallel' Package* v. R package version 1.0.15.

Microsoft Corporation and Weston, S. (2019b) *foreach: Provides Foreach Looping Construct* v. R package version 1.4.7.

Millenium Ecosystem Assessment. (2005) *Ecosystems and Human Well-Being: Wetlands and Water - Synthesis*. (World Resources Institute. Washington DC, USA).

Miller, J. and Franklin, J. (2002) Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* **157** (2–3), 227-247, doi: 10.1016/S0304-3800(02)00196-5.

Mingers, J. (1989) An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning* **4** (2), 227-243, doi: 10.1023/A:1022604100933.

Miranda, L. E. and Hodges, K. B. (2000) Role of aquatic vegetation coverage on hypoxia and sunfish abundance in bays of a eutrophic reservoir. *Hydrobiologia* **427** (1), 51-57, doi: 10.1023/A:1003999929094.

Mitsch, W. J. and Gosselink, J. G. (2000) The value of wetlands: importance of scale and landscape setting. *Ecological Economics* **35** (1), 25-33, doi: 10.1016/S0921-8009(00)00165-8.

Montgomery, W. I.; Lundy, M. G. and Reid, N. (2012) 'Invasional meltdown': evidence for unexpected consequences and cumulative impacts of multispecies invasions. *Biological Invasions* **14** (6), 1111-1125, doi: 10.1007/s10530-011-0142-4.

Mount, N. J.; Maier, H. R.; Toth, E.; Elshorbagy, A.; Solomatine, D.; Chang, F. J. and Abrahart, R. J. (2016) Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. *Hydrological Sciences Journal* **61** (7), 1192-1208, doi: 10.1080/02626667.2016.1159683.

Mouton, A. *A critical analysis of performance criteria for the evaluation and optimisation of fuzzy models for species distribution* PhD thesis, Ghent, (2008).

Mouton, A. M.; Alcaraz-Hernández, J. D.; De Baets, B.; Goethals, P. L. M. and Martínez-Capel, F. (2011) Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. *Environmental Modelling & Software* **26** (5), 615-622, doi: 10.1016/j.envsoft.2010.12.001.

Mouton, A. M.; De Baets, B. and Goethals, P. L. M. (2010) Ecological relevance of performance criteria for species distribution models. *Ecological Modelling* **221** (16), 1995-2002, doi: 10.1016/j.ecolmodel.2010.04.017.

Mouton, A. M.; Schneider, M.; Peter, A.; Holzer, G.; Müller, R.; Goethals, P. L. M. and De Pauw, N. (2008) Optimisation of a fuzzy physical habitat model for spawning European grayling (Thymallus thymallus L.) in the Aare river (Thun, Switzerland). *Ecological Modelling* **215** (1–3), 122-132, doi: 10.1016/j.ecolmodel.2008.02.028.

Mozdzer, T. J.; Zieman, J. C. and McGlathery, K. J. (2010) Nitrogen Uptake by Native and Invasive Temperate Coastal Macrophytes: Importance of Dissolved Organic Nitrogen. *Estuaries and Coasts* **33** (3), 784-797, doi: 10.1007/s12237-009-9254-9.

Muradov, N.; Taha, M.; Miranda, A. F.; Kadali, K.; Gujar, A.; Rochfort, S.; Stevenson, T.; Ball, A. S. and Mouradov, A. (2014) Dual application of duckweed and azolla plants for wastewater treatment and renewable fuels and petrochemicals production. *Biotechnol Biofuels* **7** (1), 30, doi: 10.1186/1754-6834-7-30.

Murphy, K.; Efremov, A.; Davidson, T. A.; Molina-Navarro, E.; Fidanza, K.; Crivelari Betiol, T. C.; Chambers, P.; Tapia Grimaldo, J.; Varandas Martins, S.; Springuel, I.; Kennedy, M.; Mormul, R. P.; Dibble, E.; Hofstra, D.; Lukács, B. A.; Gebler, D.; Baastrup-Spohr, L. and Urrutia-Estrada, J. (2019) World distribution, diversity and endemism of aquatic macrophytes. *Aquatic Botany* **158**, 103127, doi: 10.1016/j.aquabot.2019.06.006.

Murphy, M. A.; Evans, J. S. and Storfer, A. (2010) Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology* **91** (1), 252-261, doi: 10.1890/08-0879.1.

Myers, J. H.; Simberloff, D.; Kuris, A. M. and Carey, J. R. (2000) Eradication revisited: dealing with exotic species. *Trends in Ecology & Evolution* **15** (8), 316-320, doi: 10.1016/S0169-5347(00)01914-5.

Nelson, S. M.; Roline, R. A.; Thullen, J. S.; Sartoris, J. J. and Boutwell, J. E. (2000) Invertebrate Assemblages and Trace Element Bioaccumulation Associated with Constructed Wetlands. *Wetlands* **20** (2), 406-415, doi: 10.1672/0277-5212(2000)020[0406:IAATEB]2.0.CO;2.

Netten, J. J. C.; Arts, G. H. P.; Gylstra, R.; van Nes, E. H.; Scheffer, M. and Roijackers, R. M. M. (2010) Effect of temperature and nutrients on the competition between free-floating *Salvinia natans* and submerged *Elodea nuttallii* in mesocosms. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* **177** (2), 125-132, doi: 10.1127/1863-9135/2010/0177-0125.

Njambuya, J.; Stiers, I. and Triest, L. (2011) Competition between *Lemna minuta* and *Lemna minor* at different nutrient concentrations. *Aquatic Botany* **94** (4), 158-164, doi: 10.1016/j.aquabot.2011.02.001.

Nogueira, B. M.; Santos, T. R. A. and Zarate, L. E. in *2007 IEEE Symposium on Computational Intelligence and Data Mining*. 66-72.

O'Farrell, I.; De Tezanos Pinto, P.; Rodríguez, P. L.; Chaparro, G. and Pizarro, H. N. (2009) Experimental evidence of the dynamic effect of free-floating plants on phytoplankton ecology. *Freshwater Biology* **54** (2), 363-375, doi: 10.1111/j.1365-2427.2008.02117.x.

Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i. and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19** (16), 2088-2096, doi: 10.1093/bioinformatics/btg287.

OECD. in *Lemna sp. Growth Inhibition Test* 22 (OECD, 2006).

Olden, J. D.; Poff, N. L. and Bledsoe, B. P. (2006) Incorporating ecological knowledge into ecoinformatics: An example of modeling hierarchically structured aquatic communities with neural networks. *Ecological Informatics* **1** (1), 33-42, doi: 10.1016/j.ecoinf.2005.08.003.

Olsen, R. L.; Chappell, R. W. and Loftis, J. C. (2012) Water quality sample collection, data treatment and results presentation for principal components analysis – literature review and Illinois River watershed case study. *Water Research* **46** (9), 3110-3122, doi: 10.1016/j.watres.2012.03.028.

Ordóñez Galán, C.; Matías, J. M.; Rivas, T. and Bastante, F. G. (2009) Reforestation planning using Bayesian networks. *Environmental Modelling & Software* **24** (11), 1285-1292, doi: 10.1016/j.envsoft.2009.05.009.

Oron, G.; de-Vegt, A. and Porath, D. (1988) Nitrogen removal and conversion by duckweed grown on waste-water. *Water Research* **22** (2), 179-184, doi: 10.1016/0043-1354(88)90076-0.

Ort, C. and Siegrist, H. (2009) Assessing wastewater dilution in small rivers with high resolution conductivity probes. *Water Science and Technology* **59** (8), 1593-1601, doi: 10.2166/wst.2009.174.

Oshiro, T. M.; Perez, P. S. and Baranauskas, J. A. (2012) How Many Trees in a Random Forest? in *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings* (ed P. Perner) 154-168 (Springer Berlin Heidelberg, Berlin, Heidelberg).

Paolacci, S.; Harrison, S. and Jansen, M. A. K. (2016) A comparative study of the nutrient responses of the invasive duckweed Lemna minuta, and the native, co-generic species Lemna minor. *Aquatic Botany* **134**, 47-53, doi: 10.1016/j.aquabot.2016.07.004.

Paolacci, S.; Harrison, S. and Jansen, M. A. K. (2018) Are alien species necessarily stress sensitive? A case study on Lemna minuta and Lemna minor. *Flora* **249**, 31-39, doi: 10.1016/j.flora.2018.09.004.

Park, W. H. and Polprasert, C. (2008) Roles of oyster shells in an integrated constructed wetland system designed for P removal. *Ecological Engineering* **34** (1), 50-56, doi: 10.1016/j.ecoleng.2008.05.014.

Parker, K. A.; Springall, B. T.; Garshong, R. A.; Malachi, A. N.; Dorn, L. E.; Costa-Terryll, A.; Mathis, R. A.; Lewis, A. N.; MacCheyne, C. L.; Davis, T. T.; Rice, A. D.; Varh, N. Y.; Li, H.; Schug, M. D. and Kalcounis-Rueppell, M. C. (2019) Rapid Increases in Bat Activity and Diversity after Wetland Construction in an Urban Ecosystem. *Wetlands* **39** (4), 717-727, doi: 10.1007/s13157-018-1115-5.

Pearse-Smith, S. W. D. (2012) 'Water war' in the Mekong Basin? *Asia Pacific Viewpoint* **53** (2), 147-162, doi: 10.1111/j.1467-8373.2012.01484.x.

Pejchar, L. and Mooney, H. A. (2009) Invasive Species, ecosystem services and human well-being. *Trends in Ecology & Evolution* **24** (9), 497-504, doi: 10.1016/j.tree.2009.03.016.

Penone, C.; Davidson, A. D.; Shoemaker, K. T.; Di Marco, M.; Rondinini, C.; Brooks, T. M.; Young, B. E.; Graham, C. H. and Costa, G. C. (2014) Imputation of missing data in life-history trait datasets: which approach performs the best? *Methods in Ecology and Evolution* **5** (9), 961-970, doi: 10.1111/2041-210X.12232.

Pérez-Harguindeguy, N.; Díaz, S.; Garnier, E.; Lavorel, S.; Poorter, H.; Jaureguiberry, P.; Bret-Harte, M. S.; Cornwell, W. K.; Craine, J. M.; Gurvich, D. E.; Urcelay, C.; Veneklaas, E. J.; Reich, P. B.; Poorter, L.; Wright, I. J.; Ray, P.; Enrico, L.; Pausas, J. G.; de Vos, A. C.; Buchmann, N.; Funes, G.; Quétier, F.; Hodgson, J. G.; Thompson, K.; Morgan, H. D.; ter Steege, H.; van der Heijden, M. G. A.; Sack, L.; Blonder, B.; Poschlod, P.; Vaieretti, M. V.; Conti, G.; Staver, A. C.; Aquino, S. and Cornelissen, J. H. C. (2013) New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany* **61** (3), 167-234, doi: 10.1071/bt12225.

Perrings, C.; Williamson, M.; Barbier, E.; Delfino, D.; Dalmazzone, S.; Shogren, J.; Simmons, P. and Watkinson, A. (2002) Biological Invasion Risks and the Public Good: An Economic Perspective. *Ecology and Society* **6 (1)** (1).

Peterson, A. T.; Soberón, J. and Krishtalka, L. (2015) A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology* **15** (1), 15, doi: 10.1186/s12898-015-0046-8.

Peterson, A. T.; Stewart, A.; Mohamed, K. I. and Araújo, M. B. (2008) Shifting Global Invasive Potential of European Plants with Climate Change. *PLoS ONE* **3** (6), e2441, doi: 10.1371/journal.pone.0002441.

Phillips, S. J.; Dudík, M.; Elith, J.; Graham, C. H.; Lehmann, A.; Leathwick, J. and Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19** (1), 181-197, doi: 10.1890/07-2153.1.

Pollino, C. A.; White, A. K. and Hart, B. T. (2007) Examination of conflicts and improved strategies for the management of an endangered Eucalypt species using Bayesian networks. *Ecological Modelling* **201** (1), 37-59, doi: 10.1016/j.ecolmodel.2006.07.032.

Pulliam, H. R. (2000) On the relationship between niche and distribution. *Ecology Letters* **3** (4), 349-361, doi: 10.1046/j.1461-0248.2000.00143.x.

Pyšek, P. and Richardson, D. (2007) Traits Associated with Invasiveness in Alien Plants: Where Do we Stand? in *Biological Invasions* Vol. 193 *Ecological Studies* (ed W. Nentwig) Ch. 7, 97-125 (Springer Berlin Heidelberg).

R Core Team. (2016) *R: A language and environment for statistical computing* v. 3.3.1 (Vienna, Austria).

Rahel, F. J. and Olden, J. D. (2008) Assessing the Effects of Climate Change on Aquatic Invasive Species. *Conservation Biology* **22** (3), 521-533, doi: 10.1111/j.1523-1739.2008.00950.x.

Rai, P. K. (2009) Heavy Metal Phytoremediation from Aquatic Ecosystems with Special Reference to Macrophytes. *Critical Reviews in Environmental Science and Technology* **39** (9), 697-753, doi: 10.1080/10643380801910058.

Real, R.; Barbosa, A. M. and Vargas, J. M. (2006) Obtaining Environmental Favourability Functions from Logistic Regression. *Environmental and Ecological Statistics* **13** (2), 237-245, doi: 10.1007/s10651-005-0003-3.

Reichard, S. H. and Hamilton, C. W. (1997) Predicting Invasions of Woody Plants Introduced into North America. *Conservation Biology* **11** (1), 193-203, doi: 10.1046/j.1523-1739.1997.95473.x.

Renner, I. W.; Elith, J.; Baddeley, A.; Fithian, W.; Hastie, T.; Phillips, S. J.; Popovic, G. and Warton, D. I. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution* **6** (4), 366-379, doi: 10.1111/2041-210X.12352.

Renner, I. W. and Warton, D. I. (2013) Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics* **69** (1), 274-281, doi: 10.1111/j.1541-0420.2012.01824.x.

Richardson, D. M.; Pyšek, P.; Rejmánek, M.; Barbour, M. G.; Panetta, F. D. and West, C. J. (2000) Naturalization and invasion of alien plants: concepts and definitions. *Diversity and Distributions* **6** (2), 93-107, doi: 10.1046/j.1472-4642.2000.00083.x.

Richter, B. D.; Braun, D. P.; Mendelson, M. A. and Master, L. L. (2003) Threats to Imperiled Freshwater Fauna. *Conservation Biology* **11** (5), 1081-1093, doi: 10.1046/j.1523-1739.1997.96236.x.

Riis, T.; Olesen, B.; Clayton, J. S.; Lambertini, C.; Brix, H. and Sorrell, B. K. (2012) Growth and morphology in relation to temperature and light availability during the establishment of three invasive aquatic plant species. *Aquatic Botany* **102**, 56-64, doi: 10.1016/j.aquabot.2012.05.002.

Riley, L. A. and Dybdahl, M. F. (2015) The roles of resource availability and competition in mediating growth rates of invasive and native freshwater snails. *Freshwater Biology* **60** (7), 1308-1315, doi: 10.1111/fwb.12566.

Robson, B. J. and Clay, C. J. (2005) Local and regional macroinvertebrate diversity in the wetlands of a cleared agricultural landscape in south-western Victoria, Australia. *Aquatic Conservation: Marine and Freshwater Ecosystems* **15** (4), 403-414, doi: 10.1002/aqc.675.

Rodríguez, M. and Brisson, J. (2015) Pollutant removal efficiency of native versus exotic common reed (*Phragmites australis*) in North American treatment wetlands. *Ecological Engineering* **74**, 364-370, doi: 10.1016/j.ecoleng.2014.11.005.

Roe, Gerard H.; Baker, Marcia B. and Herla, F. (2017) Centennial glacier retreat as categorical evidence of regional climate change. *Nature Geoscience* **10** (2), 95-99, doi: 10.1038/ngeo2863.

Rokach, L. (2008) *Data mining with decision trees: theory and applications*. Vol. 69 (World scientific.).

Rooney, R. C.; Davy, C.; Gilbert, J.; Prosser, R.; Robichaud, C. and Sheedy, C. (2020) Periphyton bioconcentrates pesticides downstream of catchment dominated by agricultural land use. *Science of the Total Environment* **702**, 134472, doi: 10.1016/j.scitotenv.2019.134472.

Rooth, J. E.; Stevenson, J. C. and Cornwell, J. C. (2003) Increased sediment accretion rates following invasion byPhragmites australis: The role of litter. *Estuaries* **26** (2), 475-483, doi: 10.1007/BF02823724.

Roura-Pascual, N.; Bas, J. M.; Thuiller, W.; Hui, C.; Krug, R. M. and Brotons, L. (2009) From introduction to equilibrium: reconstructing the invasive pathways of the Argentine ant in a Mediterranean region. *Global Change Biology* **15** (9), 2101-2115, doi: 10.1111/j.1365-2486.2009.01907.x.

Rousseau, D. P.; Vanrolleghem, P. A. and De Pauw, N. (2004a) Model-based design of horizontal subsurface flow constructed treatment wetlands: a review. *Water Research* **38** (6), 1484-1493, doi: 10.1016/j.watres.2003.12.013.

Rousseau, D. P. L.; Vanrolleghem, P. A. and Pauw, N. D. (2004b) Constructed wetlands in Flanders: a performance analysis. *Ecological Engineering* **23** (3), 151-163, doi: 10.1016/j.ecoleng.2004.08.001.

RStudio Team. (2015) *RStudio: Integrated Development for R* v. 0.99.903 (RStudio, Inc., Boston, MA).

Saccá, M. L.; Barra Caracciolo, A.; Di Lenola, M. and Grenni, P. in *Soil Biological Communities and Ecosystem Resilience*. (eds M. Lukac, P. Grenni and M. Gamboni) 9-24 (Springer International Publishing).

Sakadevan, K. and Bavor, H. J. (1998) Phosphate adsorption characteristics of soils, slags and zeolite to be used as substrates in constructed wetland systems. *Water Research* **32** (2), 393-399, doi: 10.1016/S0043-1354(97)00271-6.

Sala, O. E.; Stuart Chapin, F.; Iii; Armesto, J. J.; Berlow, E.; Bloomfield, J.; Dirzo, R.; Huber-Sanwald, E.; Huenneke, L. F.; Jackson, R. B.; Kinzig, A.; Leemans, R.; Lodge, D. M.; Mooney, H. A.; Oesterheld, M. n.; Poff, N. L.; Sykes, M. T.; Walker, B. H.; Walker, M. and Wall, D. H. (2000) Global Biodiversity Scenarios for the Year 2100. *Science* **287** (5459), 1770-1774, doi: 10.1126/science.287.5459.1770.

Salski, A. (1992) Fuzzy knowledge-based models in ecological research. *Ecological Modelling* **63** (1–4), 103-112, doi: 10.1016/0304-3800(92)90064-L.

Sand-Jensen, K. and Borum, J. (1991) Interactions among phytoplankton, periphyton, and macrophytes in temperate freshwaters and estuaries. *Aquatic Botany* **41** (1), 137-175, doi: 10.1016/0304-3770(91)90042-4.

Sarkar, S. (1999) Wilderness preservation and biodiversity conservation—keeping divergent goals distinct. *Bioscience* **49** (5), 405-412, doi: 10.2307/1313633.

Sauer, J.; Domisch, S.; Nowak, C. and Haase, P. (2011) Low mountain ranges: summit traps for montane freshwater species under climate change. *Biodiversity and Conservation* **20** (13), 3133-3146, doi: 10.1007/s10531-011-0140-y.

Scheffer, M.; Bakema, A. H. and Wortelboer, F. G. (1993a) MEGAPLANT: a simulation model of the dynamics of submerged plants. *Aquatic Botany* **45** (4), 341-356, doi: 10.1016/0304-3770(93)90033-S.

Scheffer, M.; Carpenter, S.; Foley, J. A.; Folke, C. and Walker, B. (2001) Catastrophic shifts in ecosystems. *Nature* **413**, 591, doi: 10.1038/35098000.

Scheffer, M.; Hosper, S. H.; Meijer, M. L.; Moss, B. and Jeppesen, E. (1993b) Alternative equilibria in shallow lakes. *Trends in Ecology & Evolution* **8** (8), 275-279, doi: 10.1016/0169-5347(93)90254-M.

Schmitt, P.; Mandel, J. and Guedj, M. (2015) A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics* **6** (1), 2-6, doi: 10.4172/2155-6180.1000224.

Scholz, M.; Harrington, R.; Carroll, P. and Mustafa, A. (2007) The Integrated Constructed Wetlands (ICW) Concept. *Wetlands* **27** (2), 337-354.

Schrage, L. J. and Downing, J. A. (2004) Pathways of Increased Water Clarity After Fish Removal from Ventura Marsh; a Shallow, Eutrophic Wetland. *Hydrobiologia* **511** (1), 215-231, doi: 10.1023/B:HYDR.0000014065.82229.c2.

Segurado, P. and Araújo, M. B. (2004) An Evaluation of Methods for Modelling Species Distributions. *Journal of Biogeography* **31** (10), 1555-1568, doi: 10.1111/j.1365-2699.2004.01076.x.

Shah, A. D.; Bartlett, J. W.; Carpenter, J.; Nicholas, O. and Hemingway, H. (2014) Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology* **179** (6), 764-774, doi: 10.1093/aje/kwt312.

Shrive, F. M.; Stuart, H.; Quan, H. and Ghali, W. A. (2006) Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology* **6** (1), 57, doi: 10.1186/1471-2288-6-57.

Shultz, J. (2003) Bolivia: The Water War Widens. *NACLA Report on the Americas* **36** (4), 34-37, doi: 10.1080/10714839.2003.11722483.

Simberloff, D. (2006) Invasional meltdown 6 years later: important phenomenon, unfortunate metaphor, or both? *Ecology Letters* **9** (8), 912-919, doi: 10.1111/j.1461-0248.2006.00939.x.

Simberloff, D.; Keitt, B.; Will, D.; Holmes, N.; Pickett, E. and Genovesi, P. (2018) Yes we can! Exciting progress and prospects for controlling invasives on islands and beyond. *Western North American Naturalist* **78** (4), 942-958, doi: 10.3398/064.078.0431.

Simberloff, D. and Von Holle, B. (1999) Positive Interactions of Nonindigenous Species: Invasional Meltdown? *Biological Invasions* **1** (1), 21-32, doi: 10.1023/A:1010086329619.

Sinclair, S. J.; White, M. D. and Newell, G. R. (2010) How useful are species distribution models for managing biodiversity under future climates. *Ecology and Society* **15** (8).

Singh, J. and Ordoñez, I. (2016) Resource recovery from post-consumer waste: important lessons for the upcoming circular economy. *Journal of Cleaner Production* **134**, 342-353, doi: 10.1016/j.jclepro.2015.12.020.

Smith, C. S.; Howes, A. L.; Price, B. and McAlpine, C. A. (2007) Using a Bayesian belief network to predict suitable habitat of an endangered mammal – The Julia Creek dunnart (Sminthopsis douglasi). *Biological Conservation* **139** (3–4), 333-347, doi: 10.1016/j.biocon.2007.06.025.

Song, K.; Kang, H.; Zhang, L. and Mitsch, W. J. (2012) Seasonal and spatial variations of denitrification and denitrifying bacterial community structure in created riverine wetlands. *Ecological Engineering* **38** (1), 130-134, doi: 10.1016/j.ecoleng.2011.09.008.

Spieles, D. J. and Mitsch, W. J. (2000) Macroinvertebrate community structure in high- and low-nutrient constructed wetlands. *Wetlands* **20** (4), 716-729, doi: 10.1672/0277-5212(2000)020[0716:MCSIHA]2.0.CO;2.

Sprague, L. A.; Oelsner, G. P. and Argue, D. M. (2017) Challenges with secondary use of multi-source water-quality data in the United States. *Water Research* **110**, 252-261, doi: 10.1016/j.watres.2016.12.024.

Srebotnjak, T.; Carr, G.; de Sherbinin, A. and Rickwood, C. (2012) A global Water Quality Index and hot-deck imputation of missing data. *Ecological Indicators* **17**, 108-119, doi: 10.1016/j.ecolind.2011.04.023.

Stekhoven, D. J. (2013) *MissForest: Nonparametric Missing value Imputation using Random Forest*. v. 1.4.

Stekhoven, D. J. and Bühlmann, P. (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** (1), 112-118, doi: 10.1093/bioinformatics/btr597.

Stiers, I.; Crohain, N.; Josens, G. and Triest, L. (2011) Impact of three aquatic invasive species on native plants and macroinvertebrates in temperate ponds. *Biological Invasions* **13** (12), 2715-2726, doi: 10.1007/s10530-011-9942-9.

Stohlgren, T. J.; Ma, P.; Kumar, S.; Rocca, M.; Morisette, J. T.; Jarnevich, C. S. and Benson, N. (2010) Ensemble Habitat Mapping of Invasive Plant Species. *Risk Analysis* **30** (2), 224-235, doi: 10.1111/j.1539-6924.2009.01343.x.

STOWA. (2001) *Limnodata Neerlandica - De aquatisch-ecologische databank voor Nederland*. Report No. 2001-32.

Strauss, S. Y.; Webb, C. O. and Salamin, N. (2006) Exotic taxa less related to native species are more invasive. *Proceedings of the National Academy of Sciences* **103** (15), 5841, doi: 10.1073/pnas.0508073103.

Strayer, D. L. (2010) Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. *Freshwater Biology* **55** (s1), 152-174, doi: 10.1111/j.1365-2427.2009.02380.x.

Strayer, D. L.; Power, M. E.; Fagan, W. F.; Pickett, S. T. A. and Belnap, J. (2003) A Classification of Ecological Boundaries. *Bioscience* **53** (8), 723-729, doi: 10.1641/0006-3568(2003)053[0723:ACOEB]2.0.CO;2.

Strobl, C.; Boulesteix, A.-L.; Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8** (1), 25, doi: 10.1186/1471-2105-8-25.

Strobl, C.; Hothorn, T. and Zeileis, A. (2009a) Party on! – A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. *The R Journal* **1** (2), 14-17.

Strobl, C.; Malley, J. and Tutz, G. (2009b) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological methods* **14** (4), 323-348, doi: 10.1037/a0016973.

Sun, F.; Pei, H.-Y.; Hu, W.-R.; Li, X.-Q.; Ma, C.-X. and Pei, R.-T. (2013) The cell damage of Microcystis aeruginosa in PACl coagulation and floc storage processes. *Separation and Purification Technology* **115**, 123-128, doi: 10.1016/j.seppur.2013.05.004.

Sundermann, A.; Stoll, S. and Haase, P. (2011) River restoration success depends on the species pool of the immediate surroundings. *Ecological Applications* **21** (6), 1962-1971, doi: 10.1890/10-0607.1.

Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P. and Feuston, B. P. (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43** (6), 1947-1958, doi: 10.1021/ci034160g.

Svitok, M.; Hrivnák, R.; Kochjarová, J.; Oťaheľová, H. and Paľove-Balang, P. (2016) Environmental thresholds and predictors of macrophyte species richness in aquatic habitats in central Europe. *Folia Geobotanica* **51** (3), 227-238, doi: 10.1007/s12224-015-9211-2.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science* **240** (4857), 1285, doi: 10.1126/science.3287615.

Tang, Y.; Harpenslager, S. F.; van Kempen, M. M. L.; Verbaarschot, E. J. H.; Loeffen, L. M. J. M.; Roelofs, J. G. M.; Smolders, A. J. P. and Lamers, L. P. M. (2017) Aquatic macrophytes can be used for wastewater polishing but not for purification in constructed wetlands. *Biogeosciences* **14** (4), 755-766, doi: 10.5194/bg-14-755-2017.

Tanner, C. (1996) Plants for constructed wetland treatment systems - A comparison of the growth and nutrient uptake of eight emergent species. *Ecological Engineering* **7**, 59-83, doi: 10.1016/0925-8574(95)00066-6.

Thomaz, S. M. and Cunha, E. R. d. (2010) The role of macrophytes in habitat structuring in aquatic ecosystems: methods of measurement, causes and consequences on animal assemblages' composition and biodiversity. *Acta Limnologica Brasiliensia* **22**, 218-236, doi: 10.4322/actalb.02202011.

Thompson, A. and Taylor, B. N. (2008) *Guide for the Use of the International Systems of Units (SI)*. (National Institute of Standards and Technology, Gaithersburg, United States).

Thuiller, W. (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* **9** (10), 1353-1362, doi: 10.1046/j.1365-2486.2003.00666.x.

Thuiller, W.; Richardson, D. M.; Rouget, M.; Procheș, Ș. and Wilson, J. R. U. (2006) Interactions Between Environment, Species Traits, and Human Uses Describe Patterns of Plant Invasions. *Ecology* **87** (7), 1755-1769, doi: 10.1890/0012-9658(2006)87[1755:IBESTA]2.0.CO;2.

Timms, R. M. and Moss, B. (1984) Prevention of growth of potentially dense phytoplankton populations by zooplankton grazing, in the presence of zooplanktivorous fish, in a shallow wetland ecosystem. *Limnology and Oceanography* **29** (3), 472-486, doi: 10.4319/lo.1984.29.3.0472.

Tobias, V. D.; Conrad, J. L.; Mahardja, B. and Khanna, S. (2019) Impacts of water hyacinth treatment on water quality in a tidal estuarine environment. *Biological Invasions* **21** (12), 3479-3490, doi: 10.1007/s10530-019-02061-2.

Toet, S.; Huibers, L. H. F. A.; Van Logtestijn, R. S. P. and Verhoeven, J. T. A. (2003) Denitrification in the periphyton associated with plant shoots and in the sediment of a wetland system supplied with sewage treatment plant effluent. *Hydrobiologia* **501** (1), 29-44, doi: 10.1023/A:1026299017464.

Travaini-Lima, F.; Milstein, A. and Sipaúba-Tavares, L. H. (2016) Seasonal Differences in Plankton Community and Removal Efficiency of Nutrients and Organic Matter in a Subtropical Constructed Wetland. *Wetlands* **36** (5), 921-933, doi: 10.1007/s13157-016-0804-1.

Trinder, C. J.; Brooker, R. W. and Robinson, D. (2013) Plant ecology's guilty little secret: understanding the dynamics of plant competition. *Functional Ecology* **27** (4), 918-929, doi: 10.1111/1365-2435.12078.

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D. and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** (6), 520-525, doi: 10.1093/bioinformatics/17.6.520.

Truu, M.; Juhanson, J. and Truu, J. (2009) Microbial biomass, activity and community composition in constructed wetlands. *Science of the Total Environment* **407** (13), 3958-3971, doi: 10.1016/j.scitotenv.2008.11.036.

Tsoar, A.; Allouche, O.; Steinitz, O.; Rotem, D. and Kadmon, R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* **13** (4), 397-405, doi: 10.1111/j.1472-4642.2007.00346.x.

Tvedt, T. (2010) Why England and not China and India? Water systems and the history of the Industrial Revolution. *Journal of Global History* **5** (1), 29-50, doi: 10.1017/S1740022809990325.

UNESCO. (2017) *The United Nations world water development report, 2017: Wastewater: the untapped resource*. 180 Pages (UNESCO. France).

UNESCO and UN-Water. (2020) *United Nations World Water Development Report 2020: Water and Climate Change*. 219 Pages (UNESCO. Paris, France).

United Nations. (2015) *Transforming Our World: The 2030 Agenda for Sustainable Development*. (United Nations).

United Nations. (2020) *United Nations - Shaping our future together*, <www.un.org/en> (Last accessed on 21/01/2020).

van Asselen, S.; Verburg, P. H.; Vermaat, J. E. and Janse, J. H. (2013) Drivers of Wetland Conversion: a Global Meta-Analysis. *PLoS ONE* **8** (11), e81292, doi: 10.1371/journal.pone.0081292.

Van Broekhoven, E.; Adriaenssens, V.; De Baets, B. and Verdonschot, P. F. M. (2006) Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. *Ecological Modelling* **198** (1–2), 71-84, doi: 10.1016/j.ecolmodel.2006.04.006.

Van Broekhoven, E. and De Baets, B. in *Fuzzy Information Processing Society, 2006. NAFIPS 2006. Annual meeting of the North American*. 102-107.

Van Broekhoven, E. and De Baets, B. (2008) Monotone Mamdani–Assilian models under mean of maxima defuzzification. *Fuzzy Sets and Systems* **159** (21), 2819-2844, doi: 10.1016/j.fss.2008.03.014.

van der Heide, T.; Roijackers, R. M. M.; van Nes, E. H. and Peeters, E. T. H. M. (2006) A simple equation for describing the temperature dependent growth of free-floating macrophytes. *Aquatic Botany* **84** (2), 171-175, doi: 10.1016/j.aquabot.2005.09.004.

Van der Lee, G. E. M.; Van der Molen, D. T.; Van den Boogaard, H. F. P. and Van der Klis, H. (2006) Uncertainty analysis of a spatial habitat suitability model and implications for ecological management of water bodies. *Landscape Ecology* **21** (7), 1019-1032, doi: 10.1007/s10980-006-6587-7.

van Donk, E. and van de Bund, W. J. (2002) Impact of submerged macrophytes including charophytes on phyto- and zooplankton communities: allelopathy versus other mechanisms. *Aquatic Botany* **72** (3), 261-274, doi: 10.1016/S0304-3770(01)00205-4.

Van Echelpoel, W.; Boets, P. and Goethals, P. L. M. (2016) Functional Response (FR) and Relative Growth Rate (RGR) Do Not Show the Known Invasiveness of Lemna minuta (Kunth). *PLoS ONE* **11** (11), e0166132, doi: 10.1371/journal.pone.0166132.

Van Echelpoel, W.; Boets, P.; Landuyt, D.; Gobeyn, S.; Everaert, G.; Bennetsen, E.; Mouton, A. and Goethals, P. L. M. (2015) Species distribution models for sustainable ecosystem management in *Developments in Environmental Modelling* Vol. 27 (eds Y.-S. Park, S. Lek, C. Baehr and S. E. Jørgensen) Ch. 6, 115-134 (Elsevier, The Netherlands).

Van Echelpoel, W. and Goethals, P. L. M. (2018) Variable importance for sustaining macrophyte presence via random forests: data imputation and model settings. *Scientific Reports* **8** (1), 14557, doi: 10.1038/s41598-018-32966-2.

van Kleunen, M.; Weber, E. and Fischer, M. (2010) A meta-analysis of trait differences between invasive and non-invasive plant species. *Ecology Letters* **13** (2), 235-245, doi: 10.1111/j.1461-0248.2009.01418.x.

Van Landuyt, W. (2007) Herkenning van de vier in België voorkomende drijvende Lemna-soorten. *Dumortiera* **91**, 16-20.

van Puijenbroek, P. J. T. M.; Cleij, P. and Visser, H. (2014) Aggregated indices for trends in eutrophication of different types of fresh water in the Netherlands. *Ecological Indicators* **36**, 456-462, doi: 10.1016/j.ecolind.2013.08.022.

VanDerWal, J.; Shoo, L. P.; Graham, C. and Williams, S. E. (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling* **220** (4), 589-594, doi: 10.1016/j.ecolmodel.2008.11.010.

Vannevel, R. (2018) Using DPSIR and Balances to Support Water Governance. *Water* **10** (2), doi: 10.3390/w10020118.

Vanni, M. J. (2002) Nutrient Cycling by Animals in Freshwater Ecosystems. *Annual Review of Ecology and Systematics* **33** (1), 341-370, doi: 10.1146/annurev.ecolsys.33.010802.150519.

Venables, W. N. and Dichmont, C. M. (2004) GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research* **70** (2–3), 319-337, doi: 10.1016/j.fishres.2004.08.011.

Verdonschot, P. F. M.; Spears, B. M.; Feld, C. K.; Brucet, S.; Keizer-Vlek, H.; Borja, A.; Elliott, M.; Kernan, M. and Johnson, R. K. (2013) A comparative review of recovery processes in rivers, lakes, estuarine and coastal waters. *Hydrobiologia* **704** (1), 453-474, doi: 10.1007/s10750-012-1294-7.

Verdonschot, P. F. M. and van Oosten-Siedlecka, A. M. (2010) *Graadmeters aquatische natuur. Analyse gegevenskwaliteit Limnodata*.

Verhoeven, J. T. A. and Meuleman, A. F. M. (1999) Wetlands for wastewater treatment: Opportunities and limitations. *Ecological Engineering* **12** (1), 5-12, doi: 10.1016/S0925-8574(98)00050-0.

Verma, R. and Suthar, S. (2014) Synchronized urban wastewater treatment and biomass production using duckweed Lemna gibba L. *Ecological Engineering* **64** (0), 337-343, doi: 10.1016/j.ecoleng.2013.12.055.

Vezza, P.; Muñoz-Mas, R.; Martinez-Capel, F. and Mouton, A. (2015) Random forests to evaluate biotic interactions in fish distribution models. *Environmental Modelling & Software* **67**, 173-183, doi: 10.1016/j.envsoft.2015.01.005.

Vilà, M. and Weiner, J. (2004) Are invasive plant species better competitors than native plant species? – evidence from pair-wise experiments. *Oikos* **105** (2), 229-238, doi: 10.1111/j.0030-1299.2004.12682.x.

Villamagna, A. M. and Murphy, B. R. (2010) Ecological and socio-economic impacts of invasive water hyacinth (Eichhornia crassipes): a review. *Freshwater Biology* **55** (2), 282-298, doi: 10.1111/j.1365-2427.2009.02294.x.

Vincent, J. and Kirkwood, A. E. (2014) Variability of water quality, metals and phytoplankton community structure in urban stormwater ponds along a vegetation gradient. *Urban Ecosystems* **17** (3), 839-853, doi: 10.1007/s11252-014-0356-1.

Vitousek, P. M. (1990) Biological Invasions and Ecosystem Processes: Towards an Integration of Population Biology and Ecosystem Studies. *Oikos* **57** (1), 7-13, doi: 10.2307/3565731.

Vitousek, P. M.; Mooney, H. A.; Lubchenco, J. and Melillo, J. M. (1997) Human Domination of Earth's Ecosystems. *Science* **277** (5325), 494-499, doi: 10.1126/science.277.5325.494.

VMM. (2019) *Fysisch-chemische kwaliteit oppervlaktewater 2018*. Report No. D/2019/6871/013, (Vlaamse Milieumaatschappij).

Vohla, C.; Kõiv, M.; Bavor, H. J.; Chazarenc, F. and Mander, Ü. (2011) Filter materials for phosphorus removal from wastewater in treatment wetlands—A review. *Ecological Engineering* **37** (1), 70-89, doi: 10.1016/j.ecoleng.2009.08.003.

Vos, C. C.; Berry, P.; Opdam, P.; Baveco, H.; Nijhof, B.; O'Hanley, J.; Bell, C. and Kuipers, H. (2008) Adapting landscapes to climate change: examples of climate-proof ecosystem networks and priority adaptation zones. *Journal of Applied Ecology* **45** (6), 1722-1731, doi: 10.1111/j.1365-2664.2008.01569.x.

Vymazal, J. (2007) Removal of nutrients in various types of constructed wetlands. *Science of the Total Environment* **380** (1), 48-65, doi: 10.1016/j.scitotenv.2006.09.014.

Vymazal, J. (2009) The use constructed wetlands with horizontal sub-surface flow for various types of wastewater. *Ecological Engineering* **35** (1), 1-17, doi: 10.1016/j.ecoleng.2008.08.016.

Vymazal, J. (2010) Constructed Wetlands for Wastewater Treatment. *Water* **2** (3), doi: 10.3390/w2030530.

Vymazal, J. (2011a) Enhancing ecosystem services on the landscape with created, constructed and restored wetlands. *Ecological Engineering* **37** (1), 1-5, doi: 10.1016/j.ecoleng.2010.07.031.

Vymazal, J. (2011b) Plants used in constructed wetlands with horizontal subsurface flow: a review. *Hydrobiologia* **674** (1), 133-156, doi: 10.1007/s10750-011-0738-9.

Vymazal, J. (2013) Emergent plants used in free water surface constructed wetlands: A review. *Ecological Engineering* **61**, 582-592, doi: 10.1016/j.ecoleng.2013.06.023.

Waljee, A. K.; Mukherjee, A.; Singal, A. G.; Zhang, Y.; Warren, J.; Balis, U.; Marrero, J.; Zhu, J. and Higgins, P. D. R. (2013) Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* **3** (8), doi: 10.1136/bmjopen-2013-002847.

Walther, G.-R. (2010) Community and ecosystem responses to recent climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365** (1549), 2019-2024, doi: 10.1098/rstb.2010.0021.

Walther, G.-R.; Roques, A.; Hulme, P. E.; Sykes, M. T.; Pyšek, P.; Kühn, I.; Zobel, M.; Bacher, S.; Botta-Dukát, Z.; Bugmann, H.; Czúcz, B.; Dauber, J.; Hickler, T.; Jarošík, V.; Kenis, M.; Klotz, S.; Minchin, D.; Moora, M.; Nentwig, W.; Ott, J.; Panov, V. E.; Reineking, B.; Robinet, C.; Semenchenko, V.; Solarz, W.; Thuiller, W.; Vilà, M.; Vohland, K. and Settele, J. (2009) Alien species in a warmer world: risks and opportunities. *Trends in Ecology & Evolution* **24** (12), 686-693, doi: 10.1016/j.tree.2009.06.008.

Wang, M.; Zhang, D. Q.; Dong, J. W. and Tan, S. K. (2017) Constructed wetlands for wastewater treatment in cold climate — A review. *Journal of Environmental Sciences* **57**, 293-311, doi: 10.1016/j.jes.2016.12.019.

Wang, Q.; Xie, H.; Ngo, H. H.; Guo, W.; Zhang, J.; Liu, C.; Liang, S.; Hu, Z.; Yang, Z. and Zhao, C. (2016) Microbial abundance and community in subsurface flow constructed wetland microcosms: role of plant presence. *Environmental Science and Pollution Research* **23** (5), 4036-4045, doi: 10.1007/s11356-015-4286-0.

Weiss, L. C.; Pötter, L.; Steiger, A.; Kruppert, S.; Frost, U. and Tollrian, R. (2018) Rising pCO2 in Freshwater Ecosystems Has the Potential to Negatively Affect Predator-Induced Defenses in Daphnia. *Current Biology* **28** (2), 327-332.e323, doi: 10.1016/j.cub.2017.12.022.

Whigham, D. F. (1999) Ecological issues related to wetland preservation, restoration, creation and assessment. *Science of the Total Environment* **240** (1), 31-40, doi: 10.1016/S0048-9697(99)00321-6.

WHO/UNICEF. (2017) *Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines*. 116 Pages (Switzerland).

Wickham, H. (2007) Reshaping Data with the reshape Package. *Journal of Statistical Software; Vol 1, Issue 12 (2007)*, doi: 10.18637/jss.v021.i12.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing AG Switzerland.).

Wiegleb, G.; Dams, H.-U.; Byeon, W. I. and Choi, G. (2017) To What Extent Can Constructed Wetlands Enhance Biodiversity? *International Journal of Environmental Science and Development* **8** (8), 561-569, doi: 10.18178/ijesd.2017.8.8.1016.

Willems, W. *Habitat Suitability Models for the analysis and prediction of macrobenthos in the North Sea* PhD thesis, Ghent, (2010).

Williams, S. L. and Grosholz, E. D. (2008) The Invasive Species Challenge in Estuarine and Coastal Environments: Marrying Management and Science. *Estuaries and Coasts* **31** (1), 3-20, doi: 10.1007/s12237-007-9031-6.

Williamson, M. and Fitter, A. (1996) The Varying Success of Invaders. *Ecology* **77** (6), 1661-1666, doi: 10.2307/2265769.

Wilson, C. D.; Roberts, D. and Reid, N. (2011) Applying species distribution modelling to identify areas of high conservation value for endangered species: A case study using Margaritifera margaritifera (L.). *Biological Conservation* **144** (2), 821-829, doi: 10.1016/j.biocon.2010.11.014.

Worrall, P.; Peberdy, K. J. and Millett, M. C. (1997) Constructed wetlands and nature conservation. *Water Science and Technology* **35** (5), 205-213, doi: 10.1016/S0273-1223(97)00070-X.

Wu, Y.; Liu, J. and Rene, E. R. (2018) Periphytic biofilms: A promising nutrient utilization regulator in wetlands. *Bioresource Technology* **248**, 44-48, doi: 10.1016/j.biortech.2017.07.081.

WWF. (2018) *Living Planet Report - 2018: Aiming Higher*. (Gland, Switzerland).

Yu, C.; Sun, C.; Yu, L.; Zhu, M.; Xu, H.; Zhao, J.; Ma, Y. and Zhou, G. (2014) Comparative Analysis of Duckweed Cultivation with Sewage Water and SH Media for Production of Fuel Ethanol. *PLoS ONE* **9** (12), e115023.

Zadeh, L. A. (1965) Fuzzy sets. *Information and Control* **8** (3), 338-353, doi: 10.1016/S0019-9958(65)90241-X.

Zamorano, M. F.; Piccone, T. and Chimney, M. J. (2018) Effects of short-duration hydraulic pulses on the treatment performance of a periphyton-based treatment wetland. *Ecological Engineering* **111**, 69-77, doi: 10.1016/j.ecoleng.2017.11.004.

Zavaleta, E. S.; Hobbs, R. J. and Mooney, H. A. (2001) Viewing invasive species removal in a whole-ecosystem context. *Trends in Ecology & Evolution* **16** (8), 454-459, doi: 10.1016/S0169-5347(01)02194-2.

Zedler, J. B. (2003) Wetlands at your service: reducing impacts of agriculture at the watershed scale. *Frontiers in Ecology and the Environment* **1** (2), 65-72, doi: 10.1890/1540-9295(2003)001[0065:WAYSRI]2.0.CO;2.

Zedler, J. B. and Kercher, S. (2004) Causes and Consequences of Invasive Plants in Wetlands: Opportunities, Opportunists, and Outcomes. *Critical Reviews in Plant Sciences* **23** (5), 431-452, doi: 10.1080/07352680490514673.

Zedler, J. B. and Kercher, S. (2005) WETLAND RESOURCES: Status, Trends, Ecosystem Services, and Restorability. *Annual Review of Environment and Resources* **30** (1), 39-74, doi: 10.1146/annurev.energy.30.050504.144248.

Zhang, S.; Zhang, C. and Yang, Q. (2003) Data preparation for data mining. *Applied Artificial Intelligence* **17** (5-6), 375-381, doi: 10.1080/713827180.

Zhang, X.; Song, X.; Wang, H. and Zhang, H. (2008) Sequential local least squares imputation estimating missing value of microarray data. *Computers in Biology and Medicine* **38** (10), 1112-1120, doi: 10.1016/j.compbiomed.2008.08.006.

Zhang, Y.; Jeppesen, E.; Liu, X.; Qin, B.; Shi, K.; Zhou, Y.; Thomaz, S. M. and Deng, J. (2017) Global loss of aquatic vegetation in lakes. *Earth-Science Reviews* **173**, 259-265, doi: 10.1016/j.earscirev.2017.08.013.

Zhao, Y.; Fang, Y.; Jin, Y.; Huang, J.; Bao, S.; Fu, T.; He, Z.; Wang, F.; Wang, M. and Zhao, H. (2015a) Pilot-scale comparison of four duckweed strains from different genera for potential application in nutrient recovery from wastewater and valuable biomass production. *Plant Biology* **17**, 82-90, doi: 10.1111/plb.12204.

Zhao, Z.; Shi, H. J.; Wang, M. L.; Cui, L.; Yang, Z. G. and Zhao, Y. (2015b) Analysis of DNA methylation of Spirodela polyrhiza (Grater Duckweed) in response to abscisic acid using methylation-sensitive amplied polymorphism. *Russian Journal of Plant Physiology* **62** (1), 127-135, doi: 10.1134/S1021443715010197.

Zhi, W. and Ji, G. (2012) Constructed wetlands, 1991–2011: A review of research development, current trends, and future directions. *Science of the Total Environment* **441**, 19-27, doi: 10.1016/j.scitotenv.2012.09.064.

Zhong, T.; Tian, Y.-H.; Song, B.-R.; Chen, Z.-J.; Zhang, Y. and Chen, Z.-H. (2016) Effect of four wetland plants on nutrient removal and growth of eutrophic algae. *Aquatic Ecosystem Health & Management* **19** (1), 49-57, doi: 10.1080/14634988.2016.1145027.

Zhou, J. L.; Zhang, Z. L.; Banks, E.; Grover, D. and Jiang, J. Q. (2009) Pharmaceutical residues in wastewater treatment works effluents and their impact on receiving river water. *Journal of Hazardous Materials* **166** (2), 655-661, doi: 10.1016/j.jhazmat.2008.11.070.

Zimmer, K. D.; Hanson, M. A. and Butler, M. G. (2000) Factors influencing invertebrate communities in prairie wetlands: a multivariate approach. *Canadian Journal of Fisheries and Aquatic Sciences* **57** (1), 76-85, doi: 10.1139/f99-180.

Zimmer, K. D.; Hanson, M. A. and Butler, M. G. (2003) Relationships among nutrients, phytoplankton, macrophytes, and fish in prairie wetlands. *Canadian Journal of Fisheries and Aquatic Sciences* **60** (6), 721-730, doi: 10.1139/f03-060.

Zurada, J. M. (1992) *Introduction to artificial neural systems*. Vol. 8 (West publishing company St. Paul.).

Zuur, A.; Ieno, E. N.; Walker, N.; Saveliev, A. A. and Smith, G. M. (2009) *Mixed effects models and extensions in ecology with R*. (Springer.).

Zuur, A. F.; Ieno, E. N. and Elphick, C. S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1** (1), 3-14, doi: 10.1111/j.2041-210X.2009.00001.x.

APPENDICES

A

Supportive Information for Chapter 4 – Data and modelling technique

A.1 Origin of the data

Table A.1: Providers of observations collected within the Limnodata Neerlandica database.

Code	Name
AGV	Waterschap Amstel Gooi en Vecht
BWN	Bekenwerkgroep Nederland
HHD	Hoogheemraadschap van Delfland
HHN	Hoogheemraadschap Hollands Noorderkwartier
HHR	Hoogheemraadschap van Rijnland
HHS	HH van Schieland en Krimpenerwaard
HSR	Hoogheemraadschap De Stichtse Rijnlanden
KUN	Kath. Universiteit Nijmegen
PGR	Provincie Groningen
PNH	Provincie Noord-Holland
POV	Provincie Overijssel
PRF	Provinsje Fryslan
PRU	Provincie Utrecht
PSC	Piscaria
RWS	Rijkswaterstaat
STO	STORA/STOWA
WA	Waterleidingbedrijf Amsterdam
WAM	Waterschap Aa en Maas
WBD	Waterschap Brabantse Delta
WD	Waterschap de Dommel
WF	Wetterskip Fryslan
WGS	Waterschap Groot-Salland
WHA	Waterschap Hunze en Aas
WHD	Waterschap Hollandse Delta
WN	Waterschap Noorderzijlvest
WPM	Waterschap Peel en Maasvallei
WRD	Waterschap Regge en Dinkel
WRIJ	Waterschap Rijn en IJssel
WRL	Waterschap Rivierenland
WRO	Waterschap Roer en Overmaas
WRW	Waterschap Reest en Wieden
WSS	Waterschap Scheldestromen
WV	Waterschap Veluwe
WVE	Waterschap Vallei en Eem
WVV	Waterschap Velt en Vecht
WZE	Waterschap Zeeuwse Eilanden
WZV	Waterschap Zeeuws-Vlaanderen
WZZ	Waterschap Zuiderzeeland

A.2 Characterisation of the physicochemical data

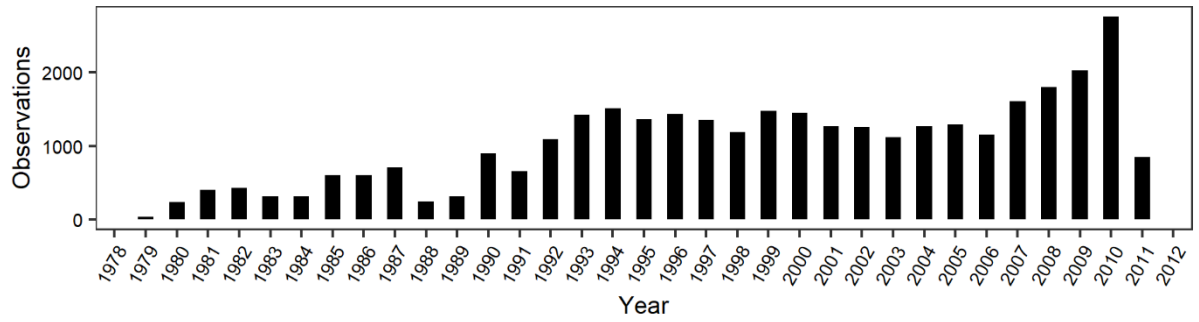

Figure A.1: Annual number of physicochemical observations. *The provided data covers a period from 1978 up to 2012, although only one observation was included for the latter. A gentle increase in observation frequency can be observed, with a maximum number of recorded observations in 2010.*

Table A.2: Overview of all 201 variables within the physicochemical data set. *Information is provided on the range, mean, median and percentage of missing data points (N_{inst} = 34 483). Variables are sorted according to increasing amount of missing data.*

Variable	Min	Max	Mean	Median	Missing (%)
Temperature	0.00	100.00	15.46	16.00	11.40
Transparency	0.00	80.00	0.57	0.40	19.67
Chloride	1.00	23000.00	412.42	96.00	21.48
Oxygen	0.00	160.00	8.82	8.90	23.30
Ammonium-nitrogen	0.00	46.00	0.38	0.20	23.72
Total phosphorus	0.00	20.00	0.40	0.19	24.95
Phosphate-phosphorus	0.00	18.00	0.22	0.05	27.78
Chlorophyll *a*	0.00	6220.00	63.79	31.00	29.81
pH (field)	2.90	78.00	7.92	8.00	34.02
Nitrite-nitrogen	0.00	60.00	0.04	0.02	38.24
Kjeldahl-nitrogen	0.01	130.00	2.40	1.90	38.69
Conductivity (field)	0.50	9060.00	162.25	63.30	42.37
Oxygen saturation	0.00	391.00	87.56	89.00	44.97
Nitrate-nitrogen	0.00	45.20	0.90	0.10	45.61
BOD5	0.00	530.00	5.54	4.00	50.03
Nitrogen oxides	0.01	45.20	1.26	0.23	57.07
Phaeophytin	0.00	1850.00	20.93	12.00	59.41
Total nitrogen	0.00	107.00	3.16	2.30	59.59
Sulphate	0.08	6200.00	99.15	60.70	59.62
Calcium	0.01	4762.67	77.33	67.00	66.81
Suspended solids	0.00	1950.00	22.69	13.70	69.14

(Continues on next page)

(Continued)

Variable	Min	Max	Mean	Median	Missing (%)
Conductivity	0.00	4100.00	131.82	77.00	76.90
Potassium	0.00	639.74	11.68	7.60	76.90
Magnesium	0.01	7800.00	34.76	11.00	77.10
Sodium	0.01	12000.00	163.00	44.00	77.89
pH	0.00	12.20	7.90	8.00	77.98
Ammonia-nitrogen	0.00	1.54	0.02	0.01	78.64
Depth	0.00	80.00	1.32	0.60	79.62
Bicarbonate	0.00	9280.43	191.95	160.00	79.82
Cupper	0.01	335.00	3.08	2.10	81.03
Zink	0.05	10000.00	20.41	10.00	81.58
Nickel	0.05	300.00	4.51	3.10	86.38
Cadmium	0.00	8.20	0.16	0.10	87.64
Iron	0.00	1600000	386.00	0.43	87.88
Lead	0.01	210.00	3.80	2.20	87.89
Chromium	0.01	79.00	2.23	1.30	88.46
Mercury	0.00	9.90	0.06	0.03	89.69
Velocity	0.00	150.00	27.99	20.00	91.05
Arsenic	0.05	143.00	4.23	2.10	93.93
Salinity	0.02	33.80	0.57	0.27	94.62
Thermo-tolerant coliforms (44 °C)	0.00	220000	2242.16	0.76	94.90
COD	2.00	466.00	59.23	60.00	95.02
Total coliforms (37 °C)	0.00	2460.00	14.30	1.00	95.52
Fluoranthene	0.00	6000.00	10.52	0.02	95.89
Benzo(a)pyrene	0.00	5000.00	7.26	0.01	95.97
Naphtalene	0.00	10000.00	14.97	0.02	96.09
Benzo(a)anthracene	0.00	5000.00	8.21	0.01	96.44
Indeno(1,2,3-c,d)pyrene	0.00	5000.00	8.26	0.01	96.46
Chrysene	0.00	5000.00	8.38	0.01	96.52
Benzo(b)fluoranthene	0.00	5000.00	8.40	0.01	96.52
Benzo(ghi)perylene	0.00	2500.00	4.39	0.01	96.66
Anthracene	0.00	5000.00	9.13	0.01	96.75
Phenanthrene	0.00	5000.00	11.72	0.02	96.84
Benzo(k)fluoranthene	0.00	900.00	1.72	0.01	96.88
Pyrene	0.00	5000.00	9.94	0.02	97.04
Dibenzo(a,h)antracene	0.00	5000.00	10.39	0.01	97.20
Fluorene	0.00	10000.00	21.28	0.01	97.26
Escherichia coli	0.10	53.52	2.05	1.00	97.59
Silica	0.04	230.00	2.29	1.10	97.71
Acenaphthylene	0.00	4.10	0.11	0.05	97.80
Acenaphthene	0.00	10.00	0.09	0.04	97.83
Alkalinity	0.04	19.80	2.76	2.70	98.05

(Continues on next page)

(Continued)

Variable	Min	Max	Mean	Median	Missing (%)
Aluminium	0.03	16400.00	350.43	107.70	98.16
alfa-Endosulphan	0.01	37.00	3.12	2.00	98.43
gamma-Hexachlorocyclohexane	0.00	260.00	5.68	4.00	98.44
Aldrin	0.01	14.00	2.86	2.00	98.44
Dieldrin	0.05	52.00	2.77	2.00	98.44
Endrin	0.10	25.00	2.61	2.00	98.44
Hexachlorobenzene	0.01	220.00	2.60	2.00	98.51
Total phosphorus (after filtration)	0.02	4.10	0.21	0.11	98.57
Zn-filtrate	0.22	300.00	9.43	5.00	98.59
beta-Endosulfan	0.00	10000.00	213.66	0.00	98.65
alpha-Hexachlorocyclohexane	0.10	13.00	3.15	1.00	98.68
beta-Hexachlorocyclohexane	0.10	43.00	4.43	5.00	98.68
Heptachlor	0.05	14.00	2.65	2.00	98.69
Heptachlor epoxide	0.02	13.00	2.33	2.00	98.69
Nickel-filtrate	0.67	30.00	3.32	2.60	98.76
Cobalt	0.20	5.00	1.18	1.00	98.90
Diazinon	4.00	600.00	31.86	20.00	98.92
Malathion	3.00	500.00	29.54	20.00	98.92
2,4-dichlorodiphenyldichloroethane	0.00	0.01	0.00	0.00	98.92
Methylparathion	3.00	800.00	31.42	10.00	98.94
Telodrin	0.00	28.00	3.09	1.00	99.00
Endosulfan sulphate	1.00	500.00	12.41	5.00	99.01
Methylazinfos	10.00	500.00	47.82	20.00	99.01
Cupper filtrate	0.50	9.00	2.52	2.00	99.02
delta-Hexachlorocyclohexane	1.00	760.00	12.66	2.00	99.03
Calcium filtrate	0.06	500000	79546.57	62000	99.03
Tin	0.20	15.00	0.79	0.20	99.07
Width	0.30	50.00	5.59	3.00	99.10
Pentachlorophenol	0.01	10.00	0.29	0.05	99.15
Lithium	0.01	17.80	0.07	0.01	99.16
Ethylazinfos	9.00	500.00	28.59	10.00	99.21
Pyrazofos	10.00	500.00	31.34	10.00	99.21
Disulfoton	3.00	500.00	25.03	10.00	99.23
Triazofos	6.00	900.00	40.54	10.00	99.26
Methyl tolclofos	0.01	7400.00	63.45	17.50	99.28

(Continues on next page)

(Continued)

Variable	Min	Max	Mean	Median	Missing (%)
Cadmium filtrate	0.00	0.70	0.17	0.20	99.28
Fenthion	2.00	160.00	10.09	10.00	99.29
Volatile organic halogenic compound	1.00	26.00	1.73	1.00	99.29
Heptenophos	3.00	610.00	21.13	9.50	99.29
Demeton	10.00	150.00	25.05	20.00	99.31
Chromium (six)	1.00	11.00	1.40	1.00	99.37
Lead-filtrate	0.10	22.00	2.11	1.00	99.44
Phenolphthalein alkalinity	0.04	0.51	0.06	0.04	99.45
Magnesium filtrate	0.01	330000	41321.44	19000	99.46
Sodium filtrate	0.03	3100000	330455.19	110000	99.47
Flow	0.00	40.00	1.14	0.05	99.55
Chromium filtrate	0.27	4.00	0.94	1.00	99.58
Mercury filtrate	0.00	0.38	0.04	0.01	99.59
Acidity	0.10	4.21	0.35	0.24	99.62
Turbidity	1.00	320.00	23.19	12.00	99.63
2,4-dichlorodifenyl dichloroethene	0.00	0.01	0.00	0.00	99.63
Iron filtrate	0.01	9.50	0.36	0.10	99.64
Atrazine	0.02	0.97	0.13	0.10	99.64
Simazine	0.10	910.00	101.33	100.00	99.65
Dimethoate	0.01	0.12	0.09	0.10	99.65
Isodrin	0.00	10.00	1.80	0.10	99.69
Ion ratio	1.78	76.00	36.46	32.94	99.70
Sum 24DDD and 44DDD	0.00	0.02	0.00	0.00	99.74
Sum 24DDE and 44DDE	0.00	0.03	0.00	0.00	99.74
Sum 24DDT and 44DDT	0.00	0.05	0.00	0.00	99.74
2,2,3,4,4,5-hexachlorobifenyl	1.00	10000.00	1094.87	2.00	99.76
2,2,4,4,5,5-hexachlorobifenyl	1.00	10000.00	1095.13	2.00	99.76
2,2,4,5,5-pentachlorobifenyl	1.00	10000.00	1094.87	2.00	99.76
2,2,5,5-tetrachlorobifenyl	0.02	200.00	13.92	2.00	99.76
2,3,4,4,5-pentachlorobifenyl	1.00	10000.00	1094.87	2.00	99.76
2,4,4-trichlorobifenyl	1.00	10000.00	1095.51	2.00	99.76
2,2,3,4,4,5,5-heptachlorobifenyl	1.00	10000.00	1115.28	2.00	99.76
Potassium filtrate	920.00	25000	9845.31	11000	99.77
Chloridazon	0.02	0.20	0.19	0.20	99.77

(Continues on next page)

(Continued)

Variable	Min	Max	Mean	Median	Missing (%)
Inorganic nitrogen	0.01	46.40	4.36	1.40	99.78
Captan	100.00	100.00	100.00	100.00	99.79
Arsenic filtrate	0.40	20.00	1.70	1.00	99.83
Benzo(b)fluorine	0.01	0.03	0.01	0.01	99.83
2,3,4,5-tetrachlorephenol	0.01	70.00	7.79	10.00	99.85
2,3,4,6-tetrachlorophenol	0.01	10.00	6.66	10.00	99.85
2,3-dichlorophenol	0.01	0.05	0.01	0.01	99.85
2,5-dichlorophenol	0.01	0.01	0.01	0.01	99.89
Diuron	20.00	1000.00	241.79	155.00	99.92
Pirimicarb	10.00	600.00	87.50	100.00	99.92
Propazin	0.01	0.50	0.15	0.10	99.92
Chlortoluron	0.02	0.21	0.06	0.05	99.93
Isoproturon	10.00	200.00	50.40	50.00	99.93
Methabenzthiazuron	20.00	60.00	32.80	30.00	99.93
Methobromuron	0.01	0.12	0.03	0.03	99.93
Metoxuron	0.02	0.30	0.04	0.03	99.93
Pentachlorobenzene	10.00	10000.00	4559.20	500.00	99.93
Aluminium filtrate	20.00	52.00	42.17	50.00	99.93
Ethoprophos	0.01	30.00	7.51	0.01	99.93
Fenitrothion	10.00	500.00	172.61	200.00	99.93
Linuron	4.00	260.00	37.62	30.00	99.94
Monolinuron	0.02	0.05	0.03	0.03	99.94
Chlorpyrifos	10.00	500.00	114.00	50.00	99.94
Terbutryn	10.00	500.00	166.50	100.00	99.94
cis-1,3-dichloropropene	0.01	1.00	0.16	0.01	99.95
Tetrachloromethane	50.00	1000.00	173.53	50.00	99.95
Trichloromethane	50.00	1000.00	197.06	50.00	99.95
1,2,3-trichloropropane	0.05	0.50	0.14	0.05	99.95
1,2-dichloropropane	0.01	0.50	0.06	0.05	99.95
1,3-dichlorobenzene	0.05	0.50	0.14	0.05	99.95
Monuron	0.01	0.07	0.03	0.03	99.95
Manganese filtrate	7.15	1200.00	156.48	70.00	99.96
Chlorbromuron	0.02	0.05	0.03	0.02	99.96
Cyanazin	0.02	1.00	0.46	0.50	99.96
Propachlor	20.00	500.00	137.14	100.00	99.96
Cobalt filtrate	0.20	1.00	0.75	1.00	99.96
Tin filtrate	0.20	0.20	0.20	0.20	99.96
Silver filtrate	1.00	5.00	2.23	1.00	99.96
Coumaphos	10.00	2000.00	773.08	1000.00	99.96
Aldicarb	50.00	1000.00	250.00	300.00	99.97
Chloroxuron	0.03	0.05	0.04	0.03	99.97
2,4-dinitrophenol	0.10	0.10	0.10	0.10	99.97

(Continues on next page)

Continued

Variable	Min	Max	Mean	Median	Missing (%)
2,5-dinitrophenol	0.10	0.10	0.10	0.10	99.97
2,6-dinitrophenol	0.10	0.10	0.10	0.10	99.97
Captafol	500.00	2000.00	1125.00	1000.00	99.98
Prometryn	200.00	500.00	300.00	250.00	99.98
1,2,3,4-tetrachlorobenzene	0.10	0.50	0.23	0.15	99.98
2,3-dichloraniline	0.20	0.50	0.33	0.30	99.98
2,4,5-trichloroaniline	0.20	0.50	0.30	0.25	99.98
2,4-dichloroaniline	0.20	0.50	0.30	0.25	99.98
2,6-dichloroaniline	0.20	1.00	0.43	0.25	99.98
Aldicarb sulfon	50.00	1000.00	208.33	50.00	99.98
Carbofuran	50.00	50.00	50.00	50.00	99.98
Hexachloroethane	0.50	0.50	0.50	0.50	99.98
Metribuzin	20.00	30.00	28.33	30.00	99.98
Oxamyl	0.05	1.30	0.36	0.20	99.98
Permethrin	20.00	20.00	20.00	20.00	99.98
Propoxur	50.00	70.00	53.33	50.00	99.98
Sum tetrachlorophenols	0.01	0.03	0.02	0.01	99.98
Sum trichlorophenols	0.01	0.06	0.03	0.02	99.98
Aldicarb sulphoxide	50.00	160.00	72.00	50.00	99.99
Carbaryl	50.00	50.00	50.00	50.00	99.99
Methomyl	50.00	140.00	74.00	50.00	99.99
Metolachlor	100.00	400.00	180.00	100.00	99.99
Bentazon	0.05	0.10	0.06	0.05	99.99
Streptococci	80.00	80.00	80.00	80.00	100.00
Desmetryn	0.01	0.01	0.01	0.01	100.00
trans-1,3-dichloropropene	1.00	1.00	1.00	1.00	100.00
2,4-dichlorophenoxy propionic acid	0.05	0.05	0.05	0.05	100.00
2,6-dichlorobenzamide	0.02	0.02	0.02	0.02	100.00

A.3 Characterisation of the macrophyte data

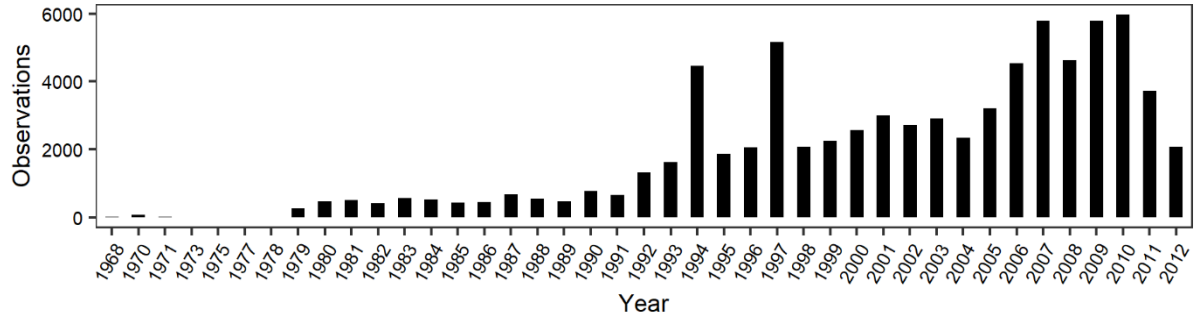

Figure A.2: Annual number of macrophyte observations. *The provided data covers a period from 1968 up to 2012, with limited observations during the first ten years. An overall increase in collection frequency can be observed, though shows a drop after reaching the maximum in 2010.*

Table A.3: Different methodologies used for macrophyte collection and identification, collected in the Limnodata Neerlandica.

Code	Method	Explanation
VEG00	Presence/Absence	No information on methodology given, simple presence/absence statements
VEG01	Tansley; water and bank	T-class
VEG02	Tansley; water	T-class, sometimes with '0' to represent presence within area (but not in sampled site)
VEG03	Tansley; bank	T-class, sometimes with '0' to represent presence within area (but not in sampled site)
VEG04	Tansley; unspecified	T-class, sometimes with '0' to represent presence within area (but not in sampled site)
VEG05	Braun-Blanquet; water	BB-class
VEG06	Braun-Blanquet; bank	BB-class
VEG07	MWTL classes	Class, 1: < 1 %; 2: 1 - 5 %; 3: 5 - 15 %; 4: 15 - 25 %; 5: 25 - 50 %; 6: 50 - 75 %; 7: > 75 %
VEG10	Coverage	Percentage, given as areal coverage per species
VEG11	Braun-Blanquet; water, modified	Class, 0: absence; 1: 3 individuals; 2: 3 individuals/m^2; 3: 4-10 individuals/m^2; 4: >10 individuals/m^2; 5-100: percentage cover per species
VEG12	Braun-Blanquet; bank, modified	Class, 0: absence; 1: 3 individuals; 2: 3 individuals/m^2; 3: 4-10 individuals/m^2; 4: >10 individuals/m^2; 5-100: percentage cover per species

(Continues on next page)

(Continued)

Code	Method	Explanation
VEG13	Attention species 1994	Class, 90/++: presence; 91/A: 1-10 individuals; 92/B: 11-25 individuals 93/C: 26-100 individuals; 94/D: 101-1000 individuals; 95/E: > 1000 individuals; 96/K: 1-10 clustered individuals; 97/L: 11-25 clustered individuals; 98/M: 26-100 clustered individuals; 99/N: 101-1000 clustered individuals; 100/P: >1000 clustered individuals; 101/V: 1-10 spread individuals; 102/W: 11-25 spread individuals; 103/X: 26-100 spread individuals; 104/Y: 101-1000 spread individuals; 105/Z: >1000 spread individuals
VEG14	Attention species 1997	Class, 90/++: presence; 91/A: 1-10 individuals; 92/B: 11-25 individuals 93/C: 26-100 individuals; 94/D: 101-1000 individuals; 95/E: > 1000 individuals; 96/K: 1-10 clustered individuals; 97/L: 11-25 clustered individuals; 98/M: 26-100 clustered individuals; 99/N: 101-1000 clustered individuals; 100/P: >1000 clustered individuals; 101/V: 1-10 spread individuals; 102/W: 11-25 spread individuals; 103/X: 26-100 spread individuals; 104/Y: 101-1000 spread individuals; 105/Z: >1000 spread individuals
VEG15	Braun-Blanquet; water and bank, unspecified	Class, 1/R: <5 % and <5 individuals; 2/+: <5 % and <3 individuals/m^2; 3/1: <5 % and 3-10 individuals/m^2; 4/2m: <5 % and >10 individuals/m^2; 5/2a: 5-12 %; 6/2b: 13-25 %; 7/3: 26-50 %; 8/4: 51-76 %; 9/5: 76-100 %
VEG16	University Nijmegen	Percentage, coverage in area of 5*5 m^2
VEG17	University Nijmegen	Percentage, coverage in area of 0.5*0.5 m^2
VEG18	Londo	Percentage
VEG19	Tansley; water, decimated	Class, 1/s: very rare; 2/r: rare or very spread; 3/o: occasionally; 4/lf: locally frequent; 5/f: frequent; 6/la: locally abundant; 7/a: high; 8/cd: co-dominant; 9/d: dominant. Mostly rooting vegetation
VEG20	Maes' range	-
VEG21	Ordinal	Class, 1: 1 %; 2: 2 %; 3: 3 %; 4: 4 %; 5: 8 %; 6: 18 %; 7: 38 %; 8: 68 %; 9: 88 %
VEG22	Water Framework Directive	T-class
VEG23	Field observations	-
VEG24	Presence	Simple presence statement
VEG25	NVO	Percentage

(Continues on next page)

(Continued)

Code	Method	Explanation
VEG26	Water Framework Directive, bank	Percentage, coverage on bank and emergent zone up to 1 m depth
VEG27	Water Framework Directive, open water	Percentage, coverage in open water
VEG28	Nat scale	Class, based on assessment in four wind directions. 1: 1 direction with limited material; 2: 1 direction with limited material; 3: 1 direction with limited material; 4: 2 directions with limited material; 5: 4 directions with limited material; 6: 1 direction with abundant material; 7: 2 directions with abundant material; 8: 3 directions with abundant material; 9: 4 directions with abundant material

A.4 Characterisation of the combined data

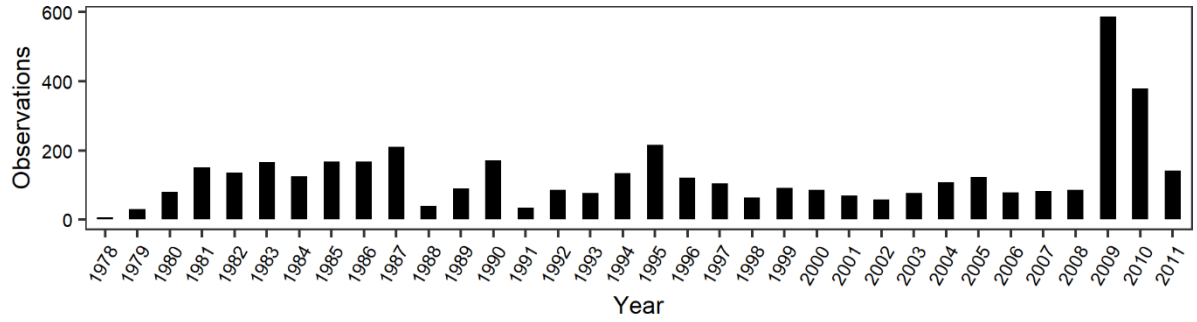

Figure A.3: Annual number of combined physicochemical and macrophyte observations. *Data contribution is spread relatively uniform among all years, except for 2009 and 2010. The temporal range is mainly determined by the availability of chemical data (see Figure A.1). Despite records for physicochemical and macrophyte observations being highest in 2010 (Figure A.1 and Figure A.2), combined information was more prevalent for 2009.*

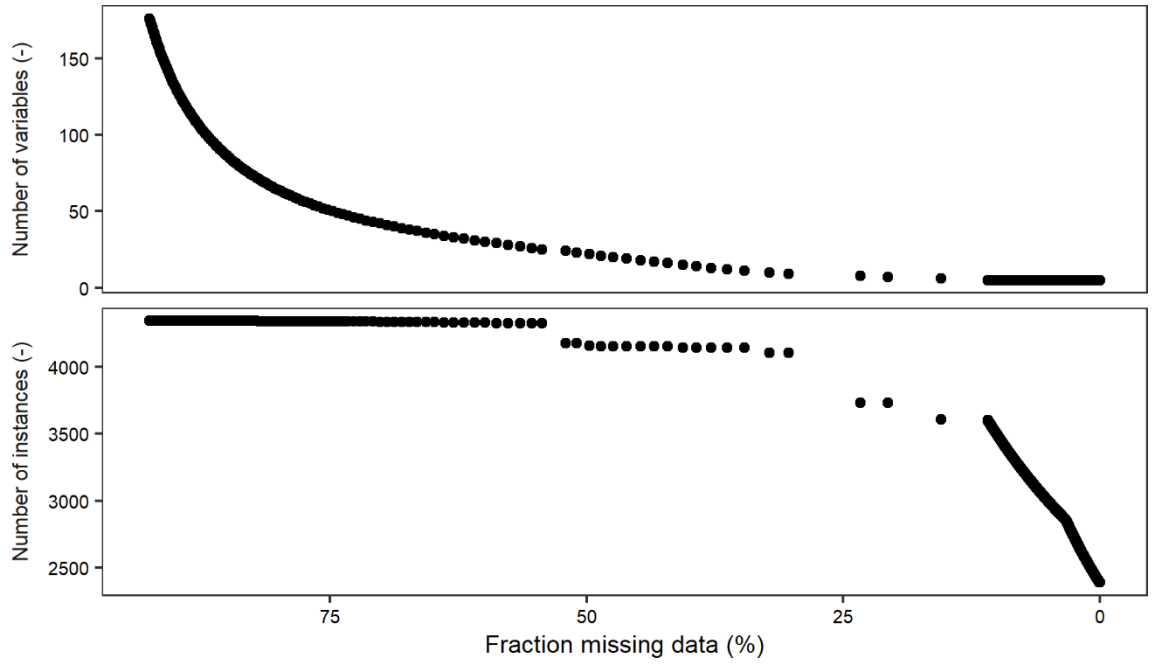

Figure A.4: Excluding variables and instances can reduce the overall degree of missing data. *Information removal was performed in a stepwise manner, removing either the variable or instance that supported the best decrease in percentage missing data.*

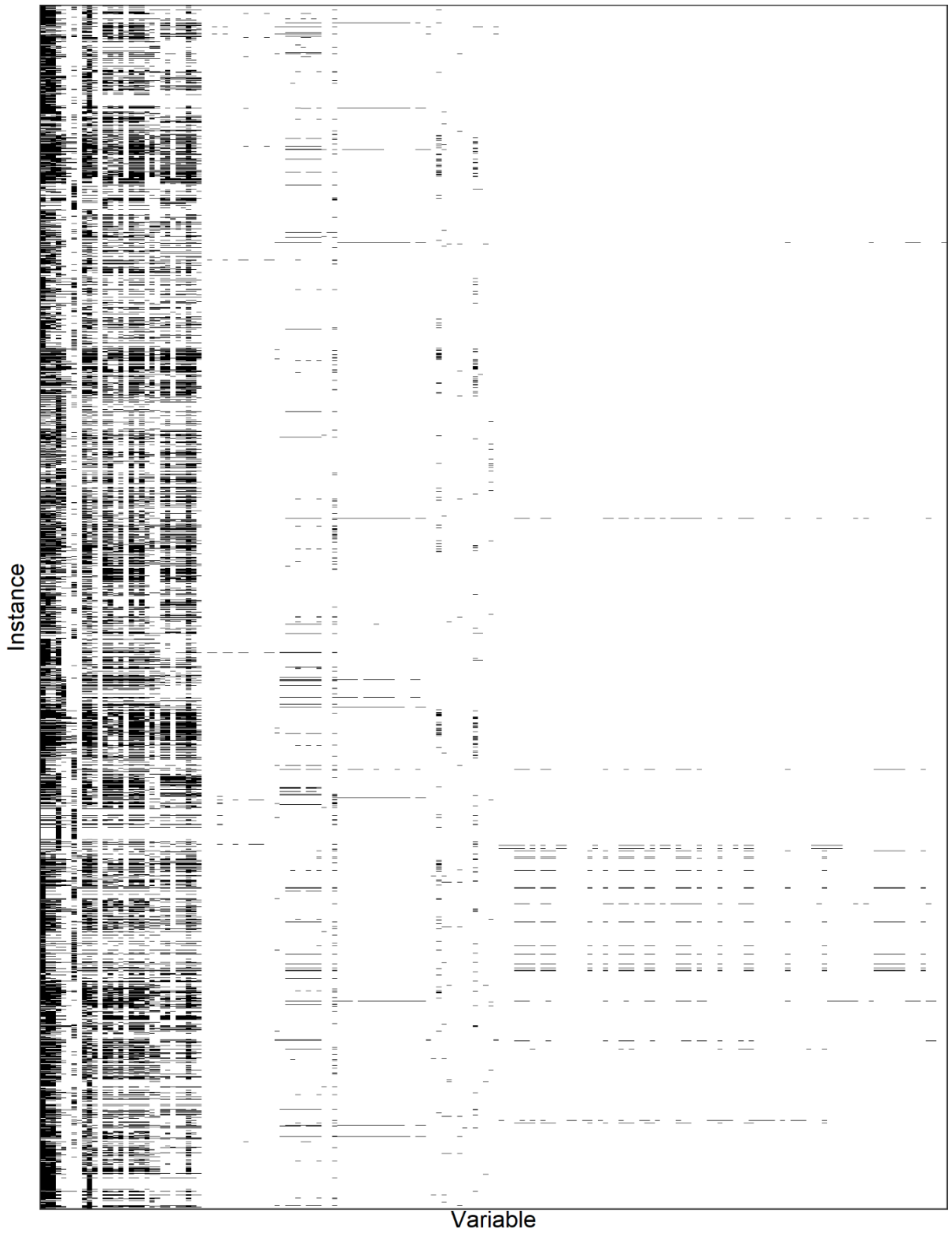

Figure A.5: Heat map of the available information within the combined data. *All instances (rows; N_{inst} = 4344) and variables (columns; N_{var} = 174) are included in this map, which indicates the presence (black) or absence (white) of a data point. It is clear that only a few variables are recorded regularly, while the majority of variables is only limitedly recorded, thereby corroborating the observations from Figure 4.5A.*

B

Supportive Information for Chapter 5 – Imputation methods for environmental data

B.1 Characterisation of the data

A detailed description of the creation of the 720 data sets is provided in Chapter 4. Construction of these data sets relies on 3 baseline data sets (see Table 4.1), which are additionally exposed to (i) random instance selection and (ii) repetitive removal of data points to obtain six different levels of missing data and ten repetitions. The variables included in these baseline data sets are mentioned in Table B.2 for data sets derived from the three baseline data sets (i.e. 5, 10 and 15 variables, Table 4.1).

Table B.1: Composition of the baseline data sets regarding number of variables and number of instances. *The first complete-case data set contained the highest number of data points. Based on this set, dimensionality for two additional data sets is pre-set during variable removal to act as baseline data (codes 2 and 3). Information is copied from Table 4.1.*

Data set code	Variable fraction (%)	Selected instances (%)	Resulting number of variables (N_{var})	Resulting number of instances (N_{inst})	Resulting number of data points
Baseline data					
1	100	100	10	17 264	172 640
2	50	100	5	21 543	107 715
3	150	100	15	3 970	59 550

Table B.2: Overview of the variables included in the baseline data sets mentioned in Table 4.1. *A distinction is made between baseline data with 5, 10 and 15 variables, representing 50 %, 100 % and 150 % of the variables within the optimal (i.e. containing most data points) data set.*

5 variables	10 variables	15 variables
Temperature	Temperature	Temperature
pH	pH	pH
Conductivity	Conductivity	Conductivity
Transparency	Transparency	Transparency
Chloride	Chloride	Chloride
	Oxygen	Oxygen
	Total phosphorus	Total phosphorus
	Phosphate-phosphorus	Phosphate-phosphorus
	Ammonium-nitrogen	Ammonium-nitrogen
	Chlorophyll *a*	Chlorophyll *a*
		Oxygen saturation
		BOD_5
		Kjeldahl-nitrogen
		Nitrite-nitrogen
		Nitrate-nitrogen

B.2 Influencing imputation performance

B.2.1 Inclusion of additional information

Including additional information has been reported to improve imputation accuracy when applying similarity-based imputation methods. As the data under consideration covers a wide range of surface water bodies (among which lakes, canals and rivers), differences in water conditions can be present, with the variance in the physicochemical data potentially being partly explained by their typology. Consequently, the inclusion of typological information was considered, but only for a subset of the data sets as part of a preliminary study.

For this specific study, each combination of data dimensionality (N_{var}) and sample size (N_{inst}) was considered for each degree of missing data (f_{MD}), resulting in a total of $3 \times 4 \times 6 = 72$ combinations (see also Table 4.1). For each combination, only the first repetition (out of 10, see Section 4.2.3.1 in Chapter 4) was considered for preliminary typology-included data imputation. The analyses were performed for *missForest* (*mF*) and *k* nearest neighbours (*kNN*), representing the similarity-based imputation methods of this study. Obtained imputation accuracies were compared with imputation accuracies of *mF* and *kNN* with default settings and without inclusion of typological information.

The results show that inclusion of typology provides similar imputation performance for *kNN*, while *mF* tends to provide lower accuracy without typological information being included in the data (see Figure B.1). Based on these observations, it was decided not to include typological information in the imputations of the other repetitions.

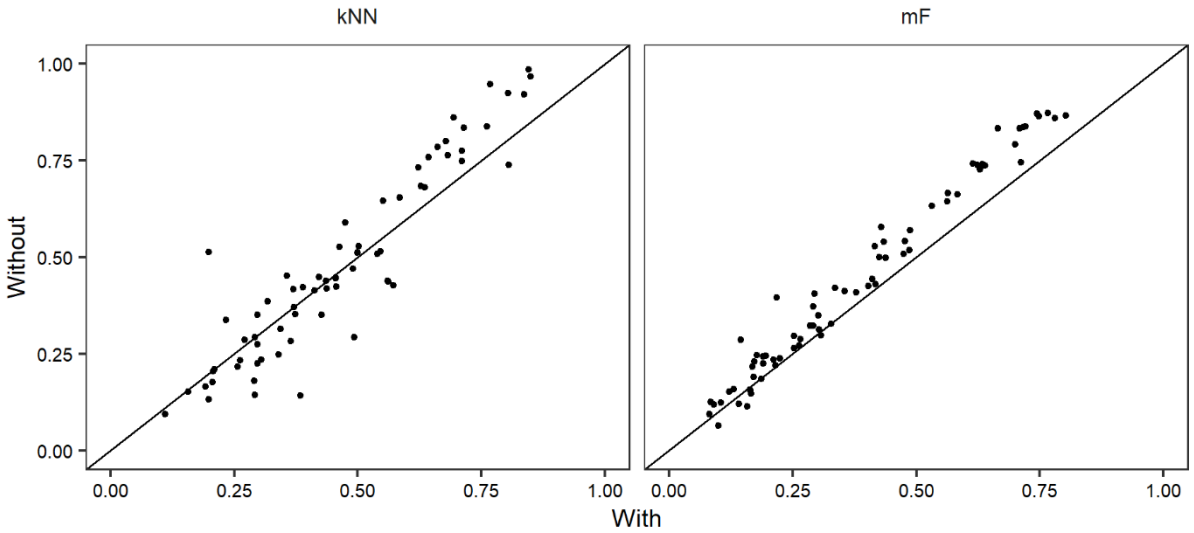

Figure B.1: Effect of including typological information during the imputation process. *Both kNN and mF show some effect of including typology on imputation accuracy. Generally, mF performs better when typological information is included (observations situated above the diagonal agreement line), while the effect on kNN accuracy is less clear.*

B.2.2 Optimisation of imputation techniques via hyperparameter setting

Of the four selected techniques, only two are characterised by a dependence on hyperparameters. More specifically, within *kNN*, the number of neighbours (k_{nn}) can be changed, while *mF* can be tuned via the number of trees (*ntree*), the number of variables selected for each split (*mtry*) and the nodesize required prior to further splitting (*nodesize*). It is expected that optimal case-specific hyperparameter settings exist and that these can be found with an iterative search, ultimately supporting improved imputation accuracy. In practice, hyperparameter optimisation started from the default setting to reduce computation time by limiting the overall search space. Hence, it remained possible that the optimised combination represented a local optimum rather than a global optimum. Implementation differed between *kNN* and *mF*, though considered every first and fifth repetition (i.e. $3 \times 4 \times 6 \times 2 = 144$ data sets) and is described in the following sections.

B.2.2.1 Nearest neighbours

The default value for k_{nn} is set to 5 within the *VIM* package. Optimisation started from this setting via a first run and calculation of performance (NRMSE). Subsequently, imputations were performed considering a range of neighbours, i.e. $k_{nn} \in [k_{nn,0} - 3, k_{nn,0} - 2, k_{nn,0} - 1, k_{nn,0} + 1, k_{nn,0} + 2, k_{nn,0} + 3]$, with $k_{nn,0}$ representing the k_{nn}-value from previous iteration, followed by re-evaluation via NRMSE. If one of the latter resulted in a lower NRMSE value, the k_{nn} value was updated and used as a new starting point. In the other case (i.e. similar performance as the previous run), new k_{nn} values were defined by extending the original range with three extra neighbours. Six extra neighbours were used if again no change in settings was observed. If the same setting was selected three times or if a total of 10 iterations was performed, the final selected settings were considered as optimal hyperparameter values.

B.2.2.2 missForest

The default settings for imputation via *missForest* are *ntree* = 100, *mtry* = $floor(\sqrt{N_{var}})$ and *nodesize* = 1. Optimisation started with the creation of three alternative starting points with *ntree* = [25, 50, 100], without changing *mtry* and *nodesize*. The settings that resulted in the lowest NRMSE value were considered for the iterative procedure. Within each iteration, the range for each settings' values was determined as follows: $[(1 - 1/(2 \cdot i)) \cdot s_o, s_o, (1 + 1/(2 \cdot i)) \cdot s_o]$, with *i* reflecting the number of iterations that resulted in the selection of the same settings and s_o reflecting the settings' value that was selected during the previous iteration. As such, the three-dimensional space of the settings' values narrows down to identify a local optimal combination. Whenever a new combination is selected, the search space is not narrowed down and simply replaces its 'central starting point'. In total, maximally ten iterations were allowed, as this showed to be sufficient to provide optimal settings' value.

B.2.3 Variability and stability among repetitions

B.2.3.1 Imputation stability (i.e. repeatability)

Imputation is reportedly case-specific and might cause different imputation results among repetitive imputation events. To test the stability of imputation, three data set combinations (cfr. Table 4.1) were selected and repetitively subjected to imputation of the missing values. More specifically, each data set was imputed three times by each method, followed by accuracy assessment via the NRMSE. The results show that the performed imputation is repeatable, with observations overlapping completely (Figure B.2).

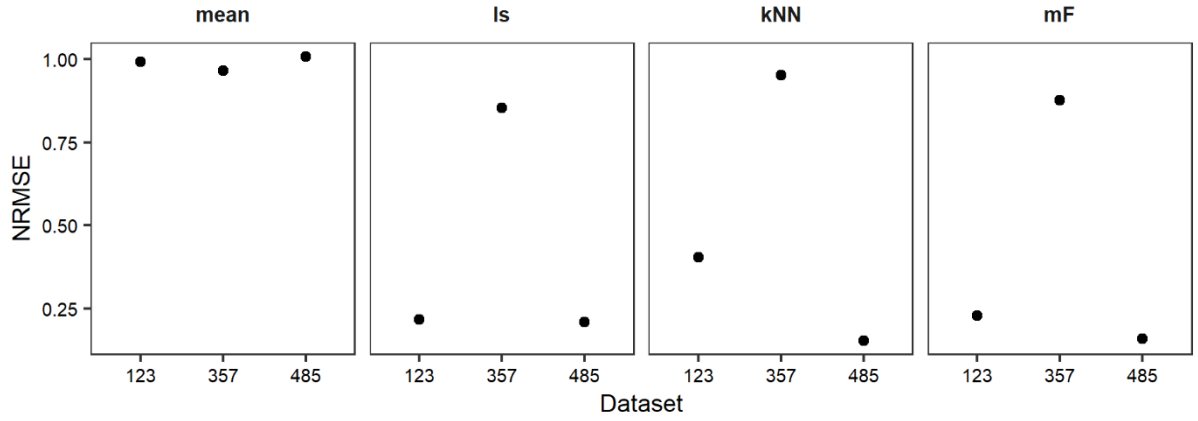

Figure B.2: Imputation stability of four imputation methods, applied thrice on three different data sets. *Complete overlap of the repetitive imputation indicates complete repeatability. The three data sets were selected randomly with ID123: 15 variables, 50 % of the instances and 1 % missing values; ID357: 10 variables, 75 % of the instances and 75 % missing values; ID485: 5 variables, 100 % of the instances and 1 % missing values. Methods: mean: mean imputation; mF: the missForest algorithm; kNN: k nearest neighbours and ls: iterative least squares. NRMSE: Normalised Root Mean Squared Error.*

B.2.3.2 Variability in optimised hyperparameters for similar combinations

Hyperparameter optimisation is case-specific, though it can be expected that similar data set characteristics support similar optimised settings. Therefore, the variability among repetitions (identical N_{var}, N_{inst} and f_{MD}, but different values being removed) is investigated. Determining the variability among ten repetitions was performed at each level of data dimensionality (5, 10 and 15 variables), both for the minimum (25 %) and maximum (100 %) sample size (cfr. Table 4.1). For each of these six combinations, the degree of missing data was set to 0.05, 0.20 or 0.75 and repeated ten times, representing an overall total of $3 \times 2 \times 3 \times 10 = 180$ data sets to be used for optimisation.

The results indicate that hyperparameter optimisation is indeed case-specific for both *mF* and *kNN*. For *mF*, the highest variability was observed for *nodesize*, though no clear pattern could be linked with the studied data set characteristics. In contrast, *mtry* was clearly negatively affected by increasing values of missing data, although this effect decreased with declining data dimensionality. In contrast, *ntree* remained relatively stable among the tested data set characteristics (Figure B.4). For *kNN*, increased dimensionality, reduced sample size and intermediate levels of missing data caused lower variability in the optimal number of neighbours (Figure B.3).

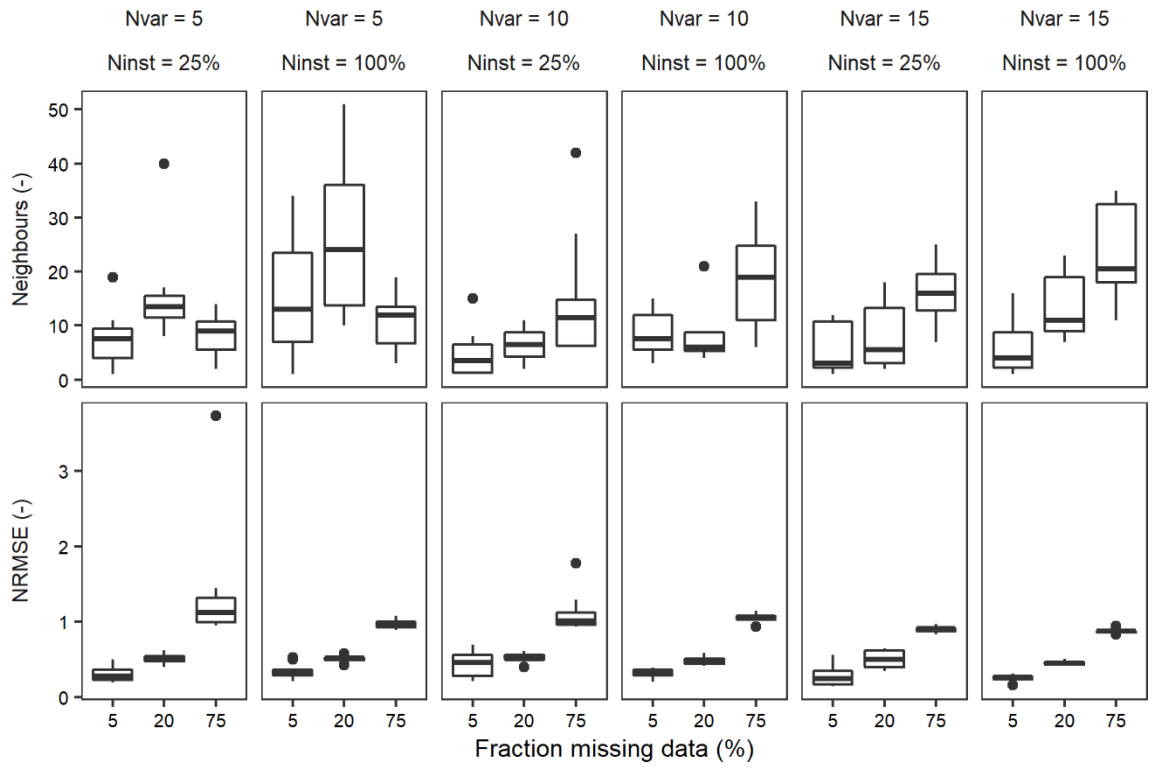

Figure B.3: Optimisation of hyperparameters of kNN, showing its case-specific character. *Optimisation was performed for ten repetitions of sample size and dimensionality, only differing in which data points were (artificially) missing. Eighteen different combinations of sample size (N_{inst}), dimensionality (N_{var}) and rate of missing data (f_{MD}) were considered. Optimised values for k_{nn} are shown along with the resulting accuracy score (NRMSE). Within the identified data set characteristics, results are separated according to rate of missing data (i.e. 5 %, 20 % and 75 %). The relative variability impedes proper value selection and highlights the case-specific properties of optimising hyperparameters. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range. Dots represent the values outside the whiskers' range.*

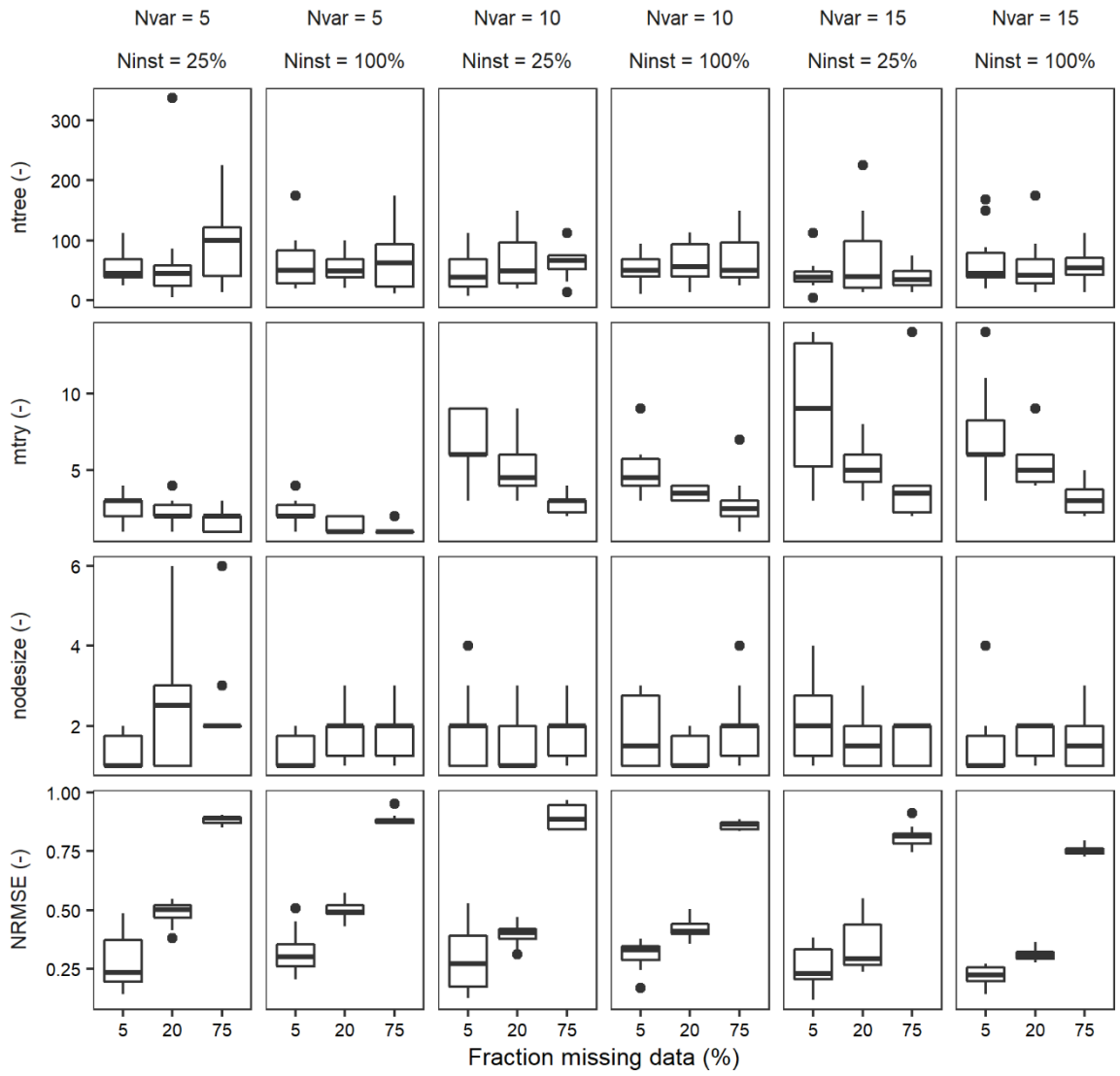

Figure B.4: Optimisation of hyperparameters of mF, showing its case-specific character. *Optimisation was performed for ten repetitions of sample size and dimensionality, only differing in which data points were (artificially) missing. Eighteen different combinations of sample size, dimensionality and rate of missing data were considered. Optimised values for ntree, mtry and nodesize are shown along with the resulting accuracy score (NRMSE). Within the identified data set characteristics, results are separated according to rate of missing data (i.e. 5 %, 20 % and 75 %). The relative variability impedes proper value selection and highlights the case-specific properties of optimising hyperparameters. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range. Dots represent the values outside the whiskers' range.*

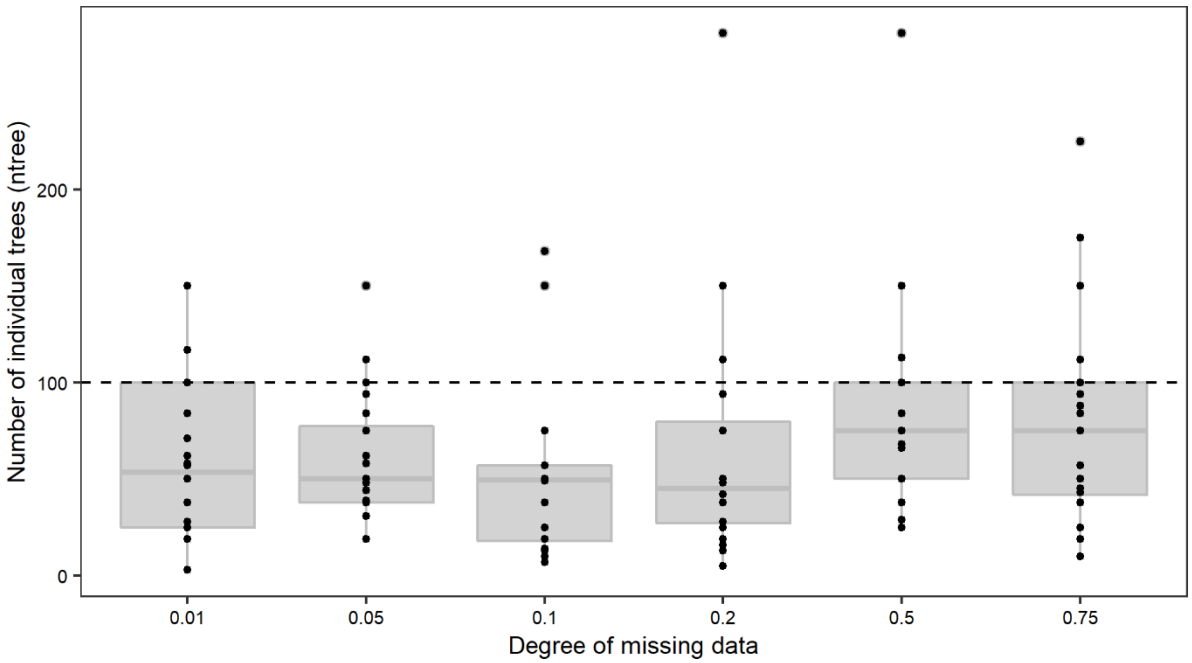

Figure B.5: Optimisation of individual trees (ntree) of mF, final values depicted according to rate of missing values. *Values range from 5 up to 225 (default: 100, represented by dashed black line) without a clear indication of a specific monotonous influence of the rate of missing values on the final ntree value.*

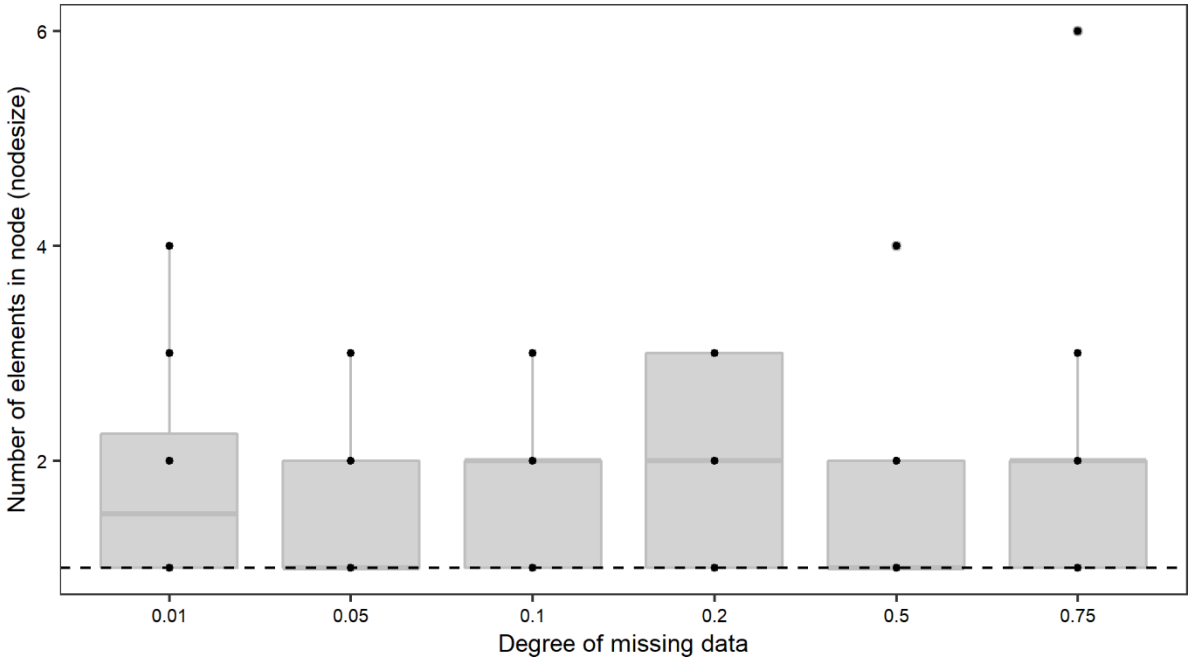

Figure B.6: Optimisation of nodesize of mF, final values depicted according to rate of missing values. *Values range from 1 up to 6 (default: 1, represented by dashed black line), with majority of data sets not requiring a clear change in nodesize to optimise accuracy.*

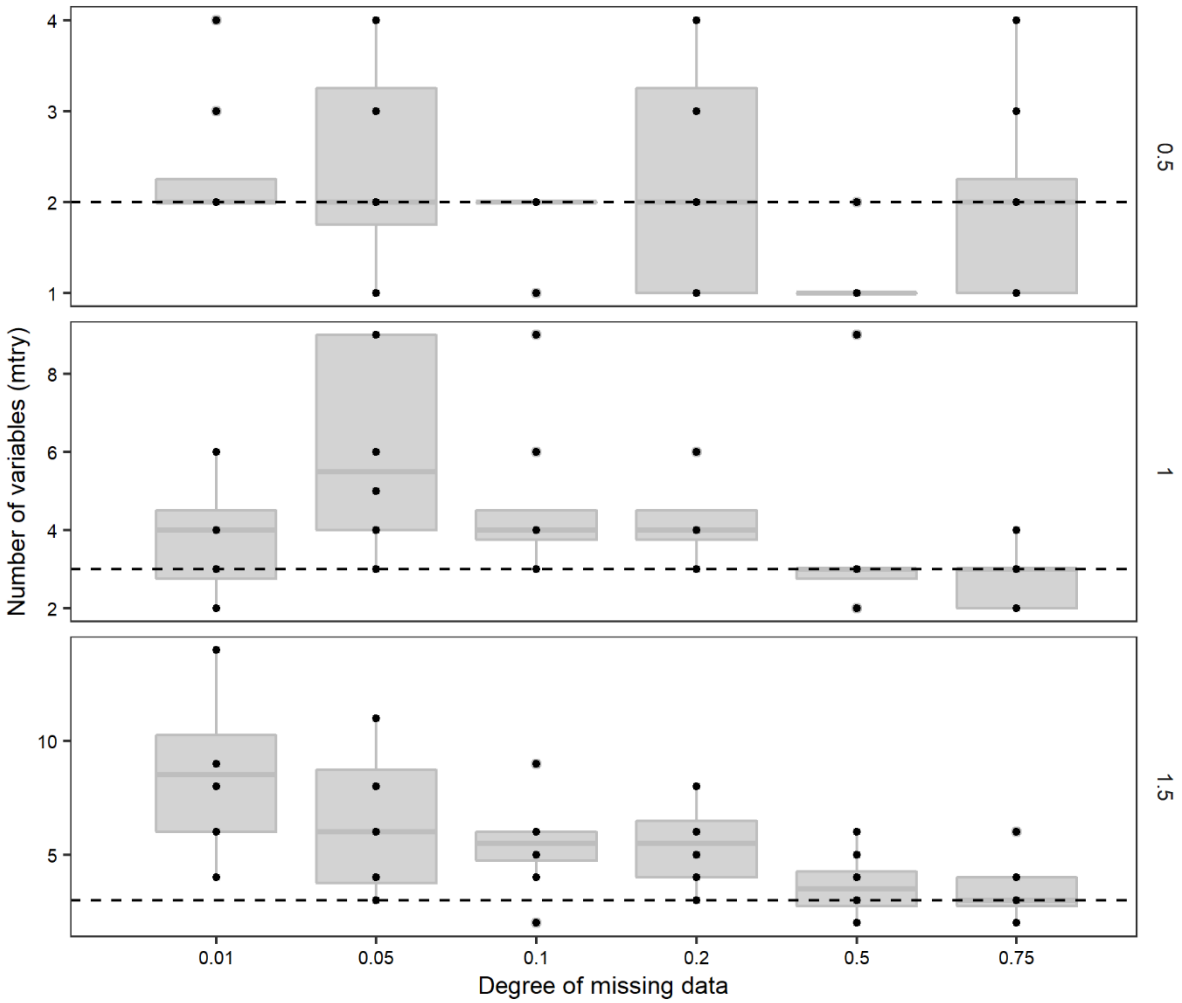

Figure B.7: Optimisation of hyperparameters of mF, final values depicted according to rate of missing values when 5 (top), 10 (middle) or 15 (bottom) variables were available. *With only 5 variables (top), values range from 1 up to 4 (default: 2, dashed black line), with majority of data requiring only 1 variable. With 10 variables (middle), values range from 1 up to 9 (default: 3, dashed black line), with majority of data requiring 3 or less variables. With 15 variables (bottom), values range from 2 up to 14 (default: 3, dashed black line), with majority of data requiring 5 or less variables. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range. Dots represent the values outside the whiskers' range.*

B.3 Results imputation performance

Table B.3: Average imputation performance for each imputation method and each combination in Table 4.1, rounded to 3 digits. *Averages are calculated based on 10 repetitions. MD: fraction missing data; Var: Relative fraction of variables considered; Obs: Fraction of instances included; ls: least squares; kNN: k nearest neighbour and mF: missForest. *Standard deviation below 0.0005, hence rounded to 0.000.*

MD	Var	Obs	Mean	ls	kNN	mF
0.01	1.5	1	0.928 ± 0.026	0.120 ± 0.053	0.133 ± 0.042	0.150 ± 0.074
0.01	1.5	0.75	0.964 ± 0.082	0.165 ± 0.128	0.238 ± 0.113	0.182 ± 0.079
0.01	1.5	0.5	1.005 ± 0.206	0.141 ± 0.104	0.275 ± 0.135	0.206 ± 0.068
0.01	1.5	0.25	1.362 ± 1.006	0.285 ± 0.532	0.325 ± 0.288	0.586 ± 1.076
0.01	1	1	0.967 ± 0.007	0.167 ± 0.048	0.195 ± 0.050	0.160 ± 0.048
0.01	1	0.75	0.969 ± 0.009	0.209 ± 0.162	0.289 ± 0.139	0.229 ± 0.166
0.01	1	0.5	0.965 ± 0.009	0.222 ± 0.189	0.284 ± 0.146	0.233 ± 0.170
0.01	1	0.25	1.024 ± 0.105	0.404 ± 0.693	0.334 ± 0.224	0.235 ± 0.177
0.01	0.5	1	0.978 ± 0.011	0.218 ± 0.137	0.244 ± 0.131	0.226 ± 0.146
0.01	0.5	0.75	0.973 ± 0.007	0.201 ± 0.146	0.227 ± 0.135	0.201 ± 0.150
0.01	0.5	0.5	0.974 ± 0.010	0.219 ± 0.197	0.204 ± 0.178	0.206 ± 0.184
0.01	0.5	0.25	0.979 ± 0.029	0.223 ± 0.081	0.238 ± 0.129	0.213 ± 0.092
0.05	1.5	1	0.927 ± 0.008	0.236 ± 0.073	0.254 ± 0.045	0.222 ± 0.042
0.05	1.5	0.75	0.934 ± 0.013	0.260 ± 0.059	0.307 ± 0.066	0.243 ± 0.037
0.05	1.5	0.5	0.927 ± 0.010	0.222 ± 0.119	0.282 ± 0.094	0.223 ± 0.063
0.05	1.5	0.25	0.937 ± 0.018	0.203 ± 0.163	0.285 ± 0.144	0.252 ± 0.088
0.05	1	1	0.966 ± 0.003	0.311 ± 0.068	0.327 ± 0.056	0.309 ± 0.063
0.05	1	0.75	0.967 ± 0.003	0.309 ± 0.200	0.324 ± 0.110	0.267 ± 0.105
0.05	1	0.5	0.965 ± 0.005	0.238 ± 0.076	0.308 ± 0.063	0.245 ± 0.074
0.05	1	0.25	0.966 ± 0.005	0.389 ± 0.351	0.438 ± 0.172	0.290 ± 0.139
0.05	0.5	1	0.976 ± 0.002	0.343 ± 0.145	0.341 ± 0.104	0.325 ± 0.096
0.05	0.5	0.75	0.974 ± 0.002	0.289 ± 0.105	0.280 ± 0.072	0.277 ± 0.068
0.05	0.5	0.5	0.975 ± 0.005	0.327 ± 0.193	0.326 ± 0.186	0.328 ± 0.148
0.05	0.5	0.25	0.974 ± 0.003	0.263 ± 0.123	0.308 ± 0.101	0.280 ± 0.123
0.1	1.5	1	0.925 ± 0.004	0.302 ± 0.065	0.338 ± 0.035	0.248 ± 0.036
0.1	1.5	0.75	0.930 ± 0.005	0.311 ± 0.054	0.369 ± 0.034	0.260 ± 0.050
0.1	1.5	0.5	0.926 ± 0.007	0.250 ± 0.072	0.314 ± 0.054	0.234 ± 0.027
0.1	1.5	0.25	0.936 ± 0.018	0.315 ± 0.171	0.388 ± 0.138	0.279 ± 0.101
0.1	1	1	0.966 ± 0.002	0.386 ± 0.096	0.394 ± 0.076	0.335 ± 0.069
0.1	1	0.75	0.967 ± 0.002	0.376 ± 0.149	0.395 ± 0.071	0.336 ± 0.062
0.1	1	0.5	0.967 ± 0.002	0.353 ± 0.126	0.378 ± 0.083	0.316 ± 0.078
0.1	1	0.25	0.967 ± 0.003	0.443 ± 0.273	0.469 ± 0.099	0.340 ± 0.094
0.1	0.5	1	0.975 ± 0.002	0.375 ± 0.092	0.380 ± 0.070	0.374 ± 0.067
0.1	0.5	0.75	0.975 ± 0.002	0.348 ± 0.083	0.360 ± 0.066	0.358 ± 0.059
0.1	0.5	0.5	0.975 ± 0.003	0.367 ± 0.104	0.381 ± 0.086	0.369 ± 0.084
0.1	0.5	0.25	0.974 ± 0.003	0.338 ± 0.068	0.354 ± 0.062	0.357 ± 0.056

(Continues on next page)

(Continued)

MD	Var	Obs	Mean	ls	kNN	mF
0.2	1.5	1	0.925 ± 0.002	0.400 ± 0.022	0.452 ± 0.027	0.310 ± 0.025
0.2	1.5	0.75	0.927 ± 0.003	0.423 ± 0.050	0.457 ± 0.049	0.329 ± 0.052
0.2	1.5	0.5	0.928 ± 0.006	0.372 ± 0.069	0.428 ± 0.036	0.302 ± 0.039
0.2	1.5	0.25	0.936 ± 0.010	0.441 ± 0.114	0.510 ± 0.116	0.347 ± 0.114
0.2	1	1	0.966 ± 0.002	0.469 ± 0.062	0.488 ± 0.058	0.424 ± 0.046
0.2	1	0.75	0.967 ± 0.002	0.461 ± 0.107	0.493 ± 0.060	0.400 ± 0.058
0.2	1	0.5	0.967 ± 0.001	0.460 ± 0.048	0.501 ± 0.026	0.420 ± 0.030
0.2	1	0.25	0.966 ± 0.002	0.422 ± 0.101	0.518 ± 0.063	0.396 ± 0.047
0.2	0.5	1	0.975 ± 0.001	0.497 ± 0.048	0.507 ± 0.047	0.497 ± 0.041
0.2	0.5	0.75	0.974 ± 0.001	0.519 ± 0.070	0.528 ± 0.056	0.507 ± 0.057
0.2	0.5	0.5	0.975 ± 0.002	0.454 ± 0.074	0.480 ± 0.057	0.475 ± 0.054
0.2	0.5	0.25	0.974 ± 0.002	0.475 ± 0.078	0.508 ± 0.069	0.487 ± 0.054
0.5	1.5	1	0.925 ± 0.001	0.605 ± 0.021	0.652 ± 0.024	0.511 ± 0.024
0.5	1.5	0.75	0.929 ± 0.002	0.609 ± 0.017	0.669 ± 0.016	0.514 ± 0.018
0.5	1.5	0.5	0.928 ± 0.001	0.597 ± 0.021	0.659 ± 0.019	0.518 ± 0.036
0.5	1.5	0.25	0.937 ± 0.006	0.620 ± 0.049	0.688 ± 0.041	0.573 ± 0.042
0.5	1	1	0.966 ± 0.001	0.692 ± 0.023	0.786 ± 0.030	0.657 ± 0.027
0.5	1	0.75	0.967 ± 0.000*	0.683 ± 0.042	0.766 ± 0.027	0.654 ± 0.035
0.5	1	0.5	0.967 ± 0.001	0.707 ± 0.036	0.767 ± 0.025	0.664 ± 0.019
0.5	1	0.25	0.967 ± 0.001	0.698 ± 0.038	0.773 ± 0.046	0.678 ± 0.044
0.5	0.5	1	0.975 ± 0.000*	0.721 ± 0.034	0.769 ± 0.036	0.739 ± 0.030
0.5	0.5	0.75	0.974 ± 0.000*	0.721 ± 0.025	0.769 ± 0.024	0.743 ± 0.034
0.5	0.5	0.5	0.974 ± 0.001	0.723 ± 0.029	0.775 ± 0.031	0.749 ± 0.034
0.5	0.5	0.25	0.974 ± 0.001	0.699 ± 0.056	0.779 ± 0.072	0.729 ± 0.044
0.75	1.5	1	0.926 ± 0.001	0.769 ± 0.017	0.878 ± 0.031	0.756 ± 0.024
0.75	1.5	0.75	0.929 ± 0.001	0.774 ± 0.015	0.847 ± 0.022	0.758 ± 0.024
0.75	1.5	0.5	0.929 ± 0.001	0.762 ± 0.020	0.863 ± 0.028	0.776 ± 0.036
0.75	1.5	0.25	0.937 ± 0.007	0.811 ± 0.031	0.903 ± 0.041	0.813 ± 0.047
0.75	1	1	0.966 ± 0.000*	0.838 ± 0.015	1.052 ± 0.055	0.860 ± 0.017
0.75	1	0.75	0.967 ± 0.000*	0.835 ± 0.024	1.080 ± 0.136	0.857 ± 0.019
0.75	1	0.5	0.967 ± 0.001	0.848 ± 0.014	0.997 ± 0.055	0.871 ± 0.032
0.75	1	0.25	0.967 ± 0.000*	0.870 ± 0.021	1.107 ± 0.259	0.895 ± 0.054
0.75	0.5	1	0.975 ± 0.000*	0.869 ± 0.017	0.977 ± 0.059	0.885 ± 0.026
0.75	0.5	0.75	0.974 ± 0.000*	0.856 ± 0.017	0.967 ± 0.055	0.960 ± 0.236
0.75	0.5	0.5	0.975 ± 0.000*	0.856 ± 0.033	1.073 ± 0.230	0.878 ± 0.022
0.75	0.5	0.25	0.974 ± 0.001	0.854 ± 0.026	1.400 ± 0.835	0.884 ± 0.018

B.4 Case studies

The analyses performed throughout Chapter 5 focused on a single metric to describe the imputation performance of each technique applied on a range of data sets. Naturally, such an aggregation causes a loss of information and limits technique-related accuracy transparency. More specifically, high errors for a single variable can inflate the normalised root mean squared error (NRMSE), which can be avoided by predictor selection or transformation. To illustrate this variable-specific imputation accuracy, two data set were selected for a more in-depth analysis.

B.4.1 Case 1: Small data set with low degree of missing data

Both brevity and visualisation were considered during the selection of the first data set and steered the decision towards a data set containing 5 variables and 5385 instances (i.e. combination 9, Table 4.1), with 1 % missing data. Hence, in total 269 data points were artificially removed prior to imputation assessment. The variables within the data were chloride ($mg \cdot L^{-1}$), conductivity ($mS \cdot m^{-1}$), pH (-), temperature (°C) and transparency (m).

B.4.1.1 Imputed values

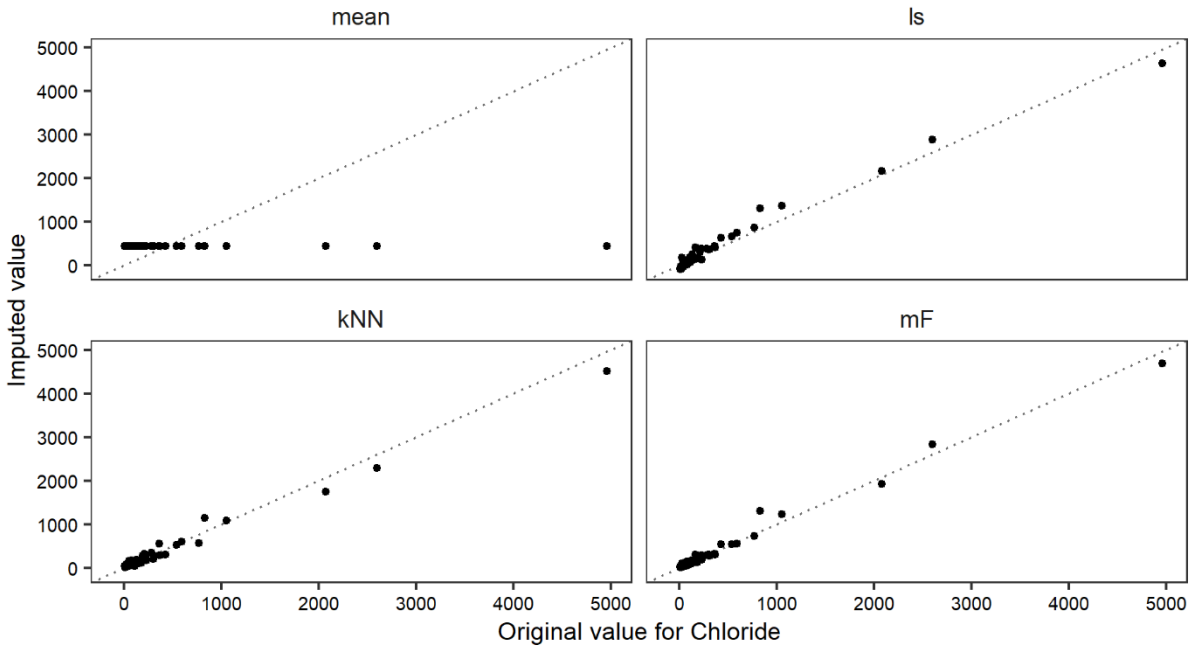

Figure B.8: Imputation of chloride by four imputation techniques. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values). Units are $mg \cdot L^{-1}$.*

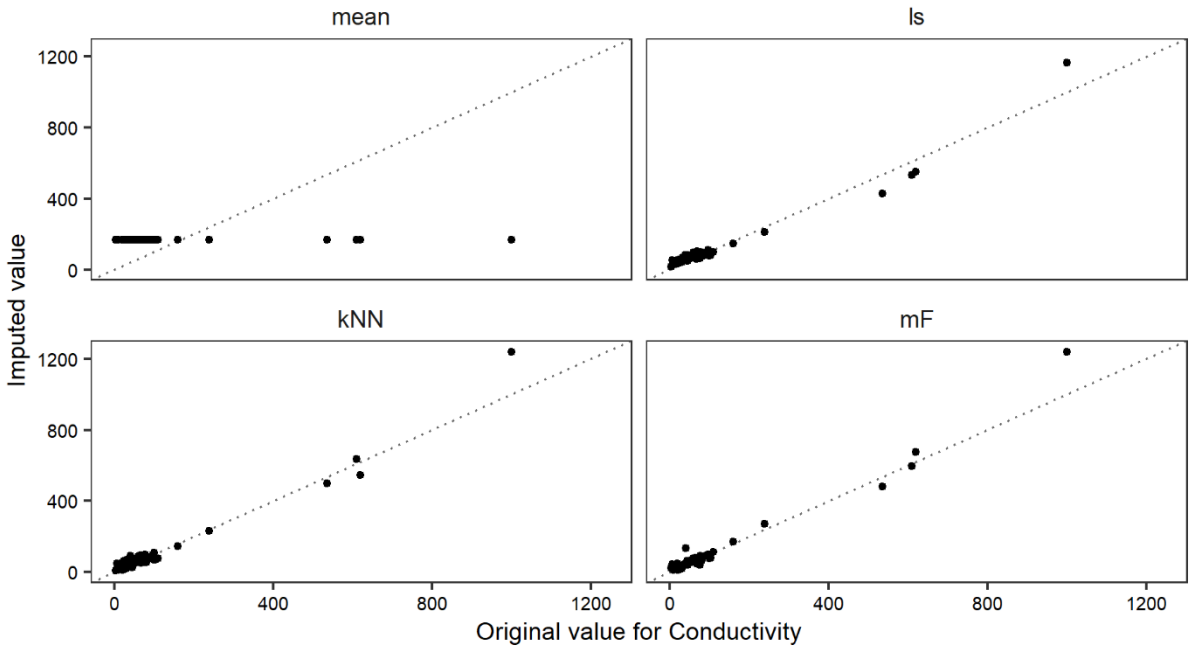

Figure B.9: Imputation of conductivity by four imputation techniques. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values). Units are mS·m⁻¹.*

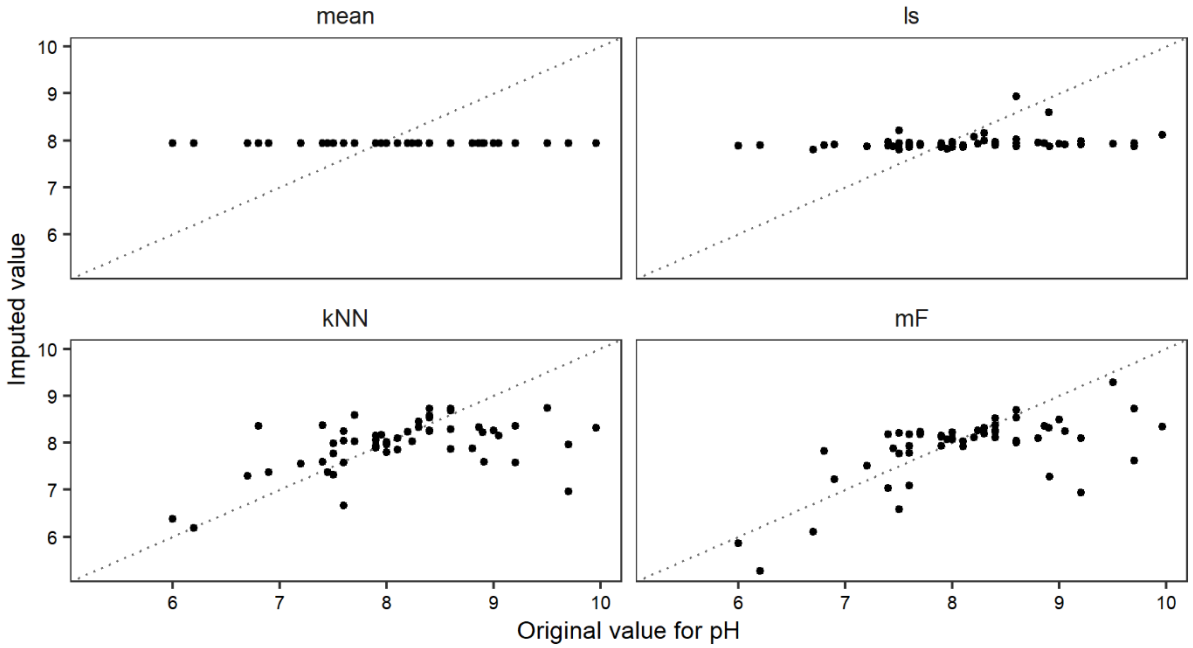

Figure B.10: Imputation of pH by four imputation techniques. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values).*

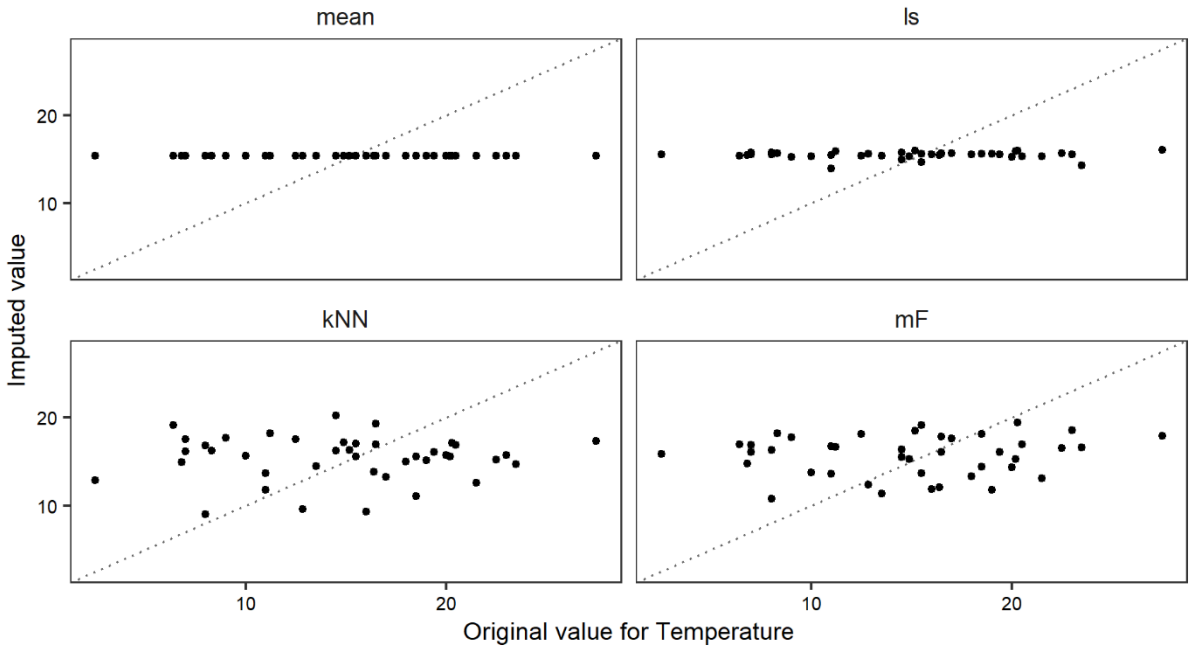

Figure B.11: Imputation of temperature by four imputation techniques. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values). Units are °C.*

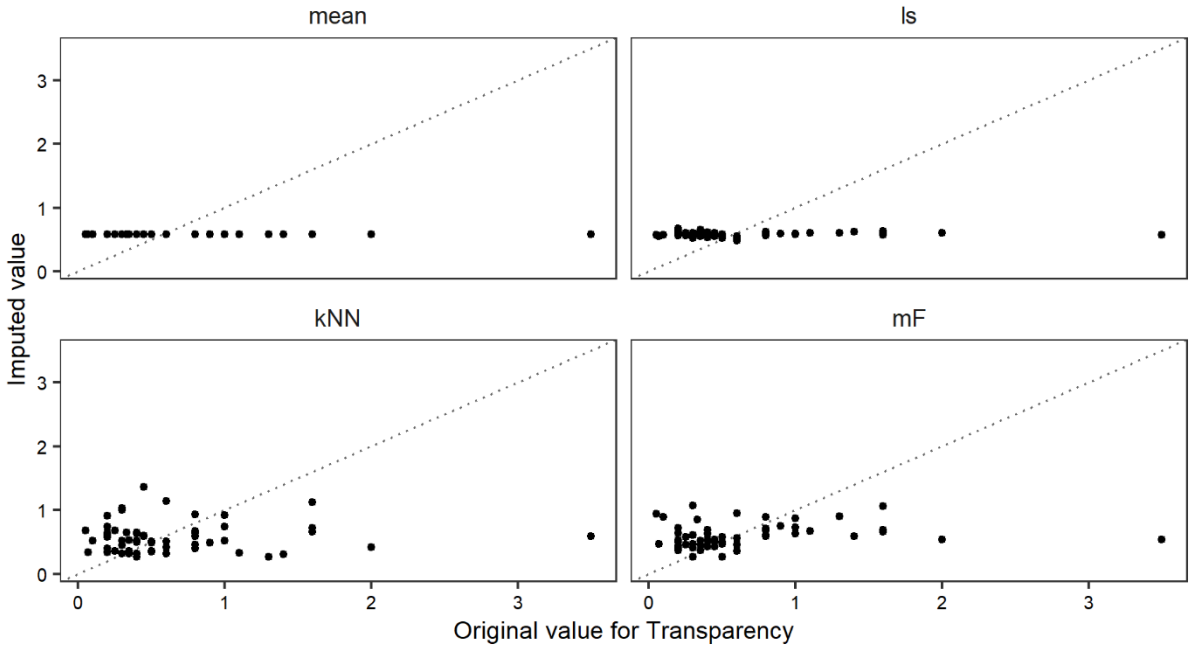

Figure B.12: Imputation of transparency by four imputation techniques. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values). Units are m.*

B.4.1.2 Variable distributions

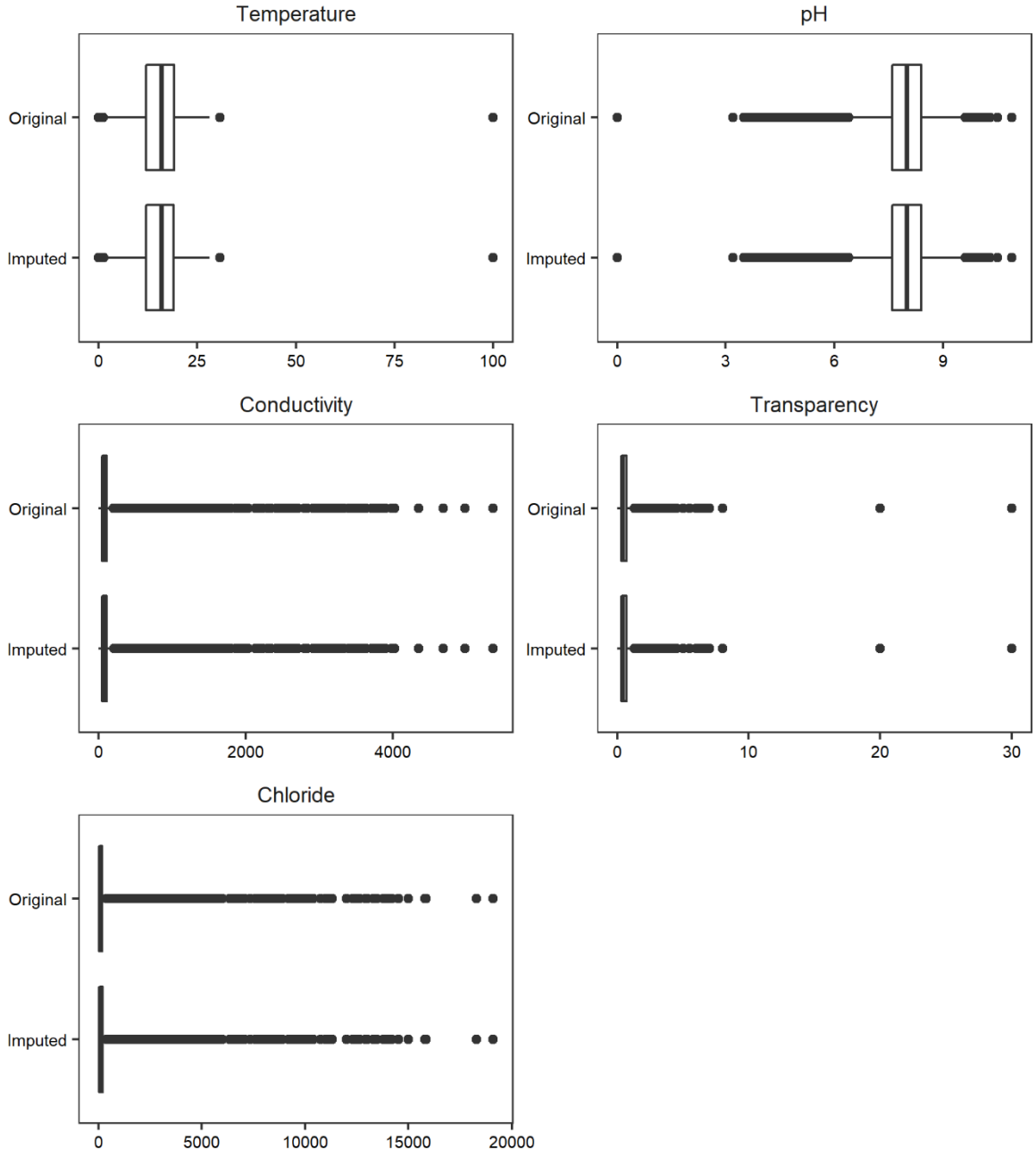

Figure B.13: Variable distributions before and after imputation by the mean. *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values).*

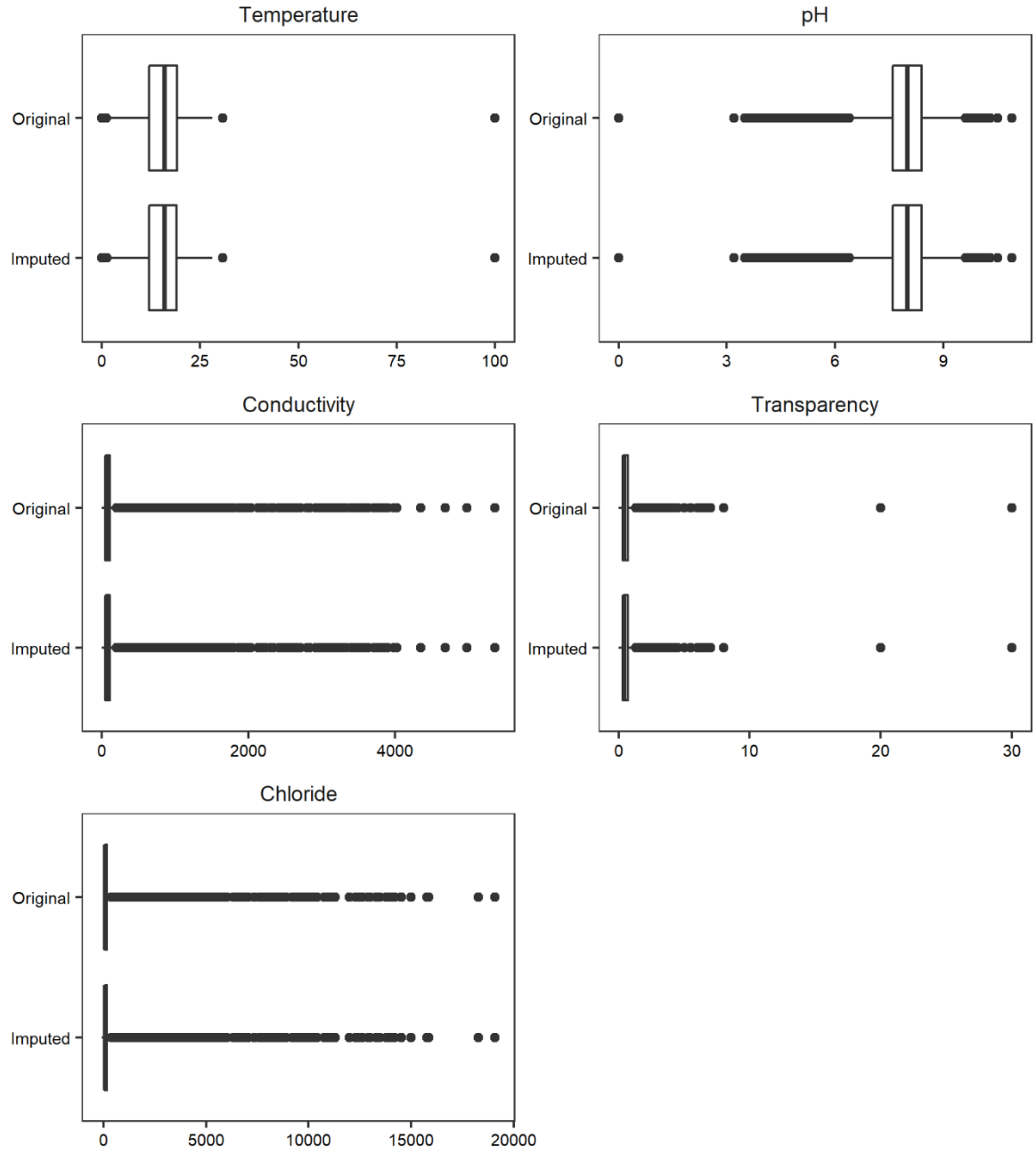

Figure B.14: Variable distributions before and after imputation by least squares regression (ls). *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values).*

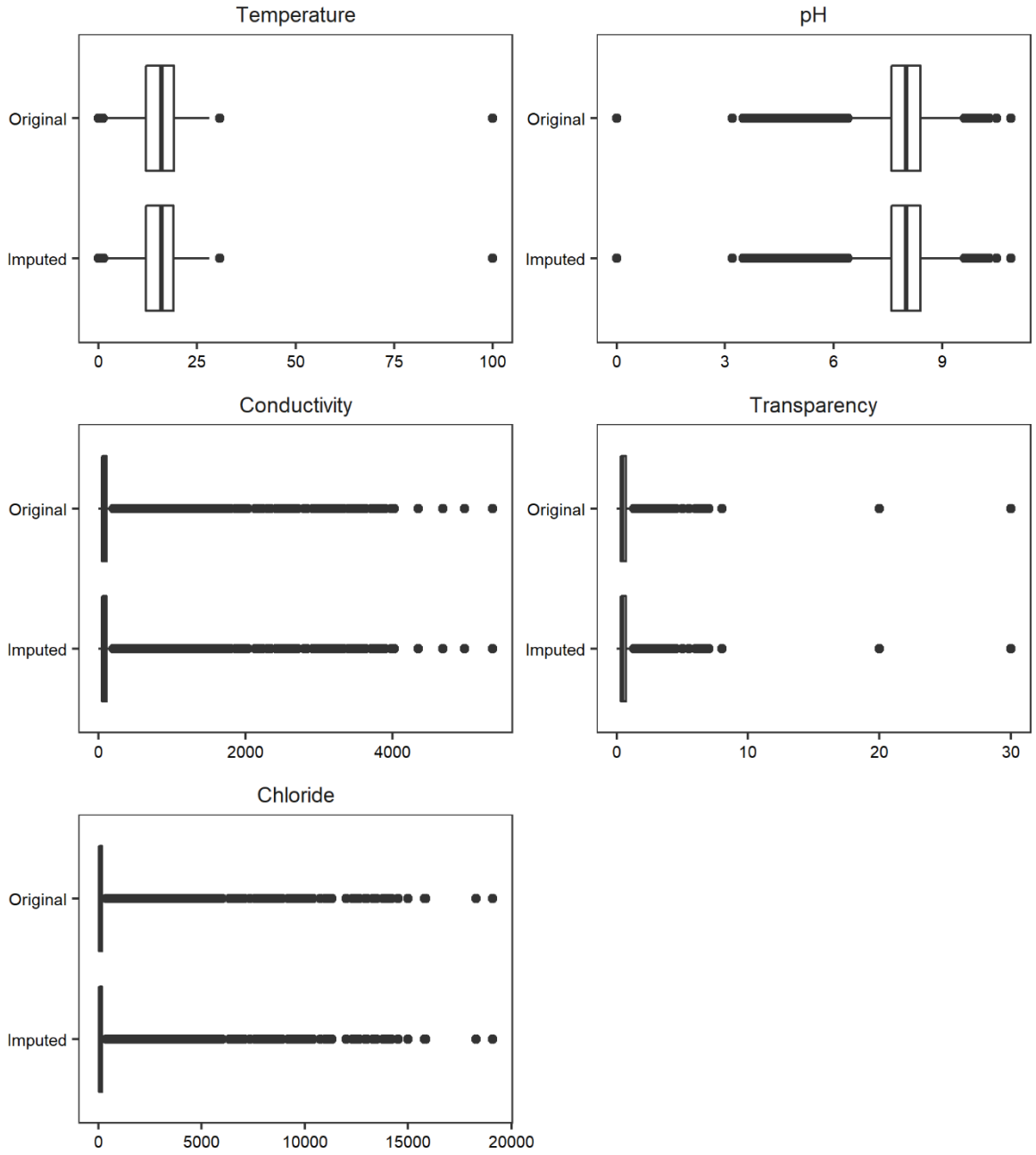

Figure B.15: Variable distributions before and after imputation by* k *nearest neighbours (*kNN*). *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values).*

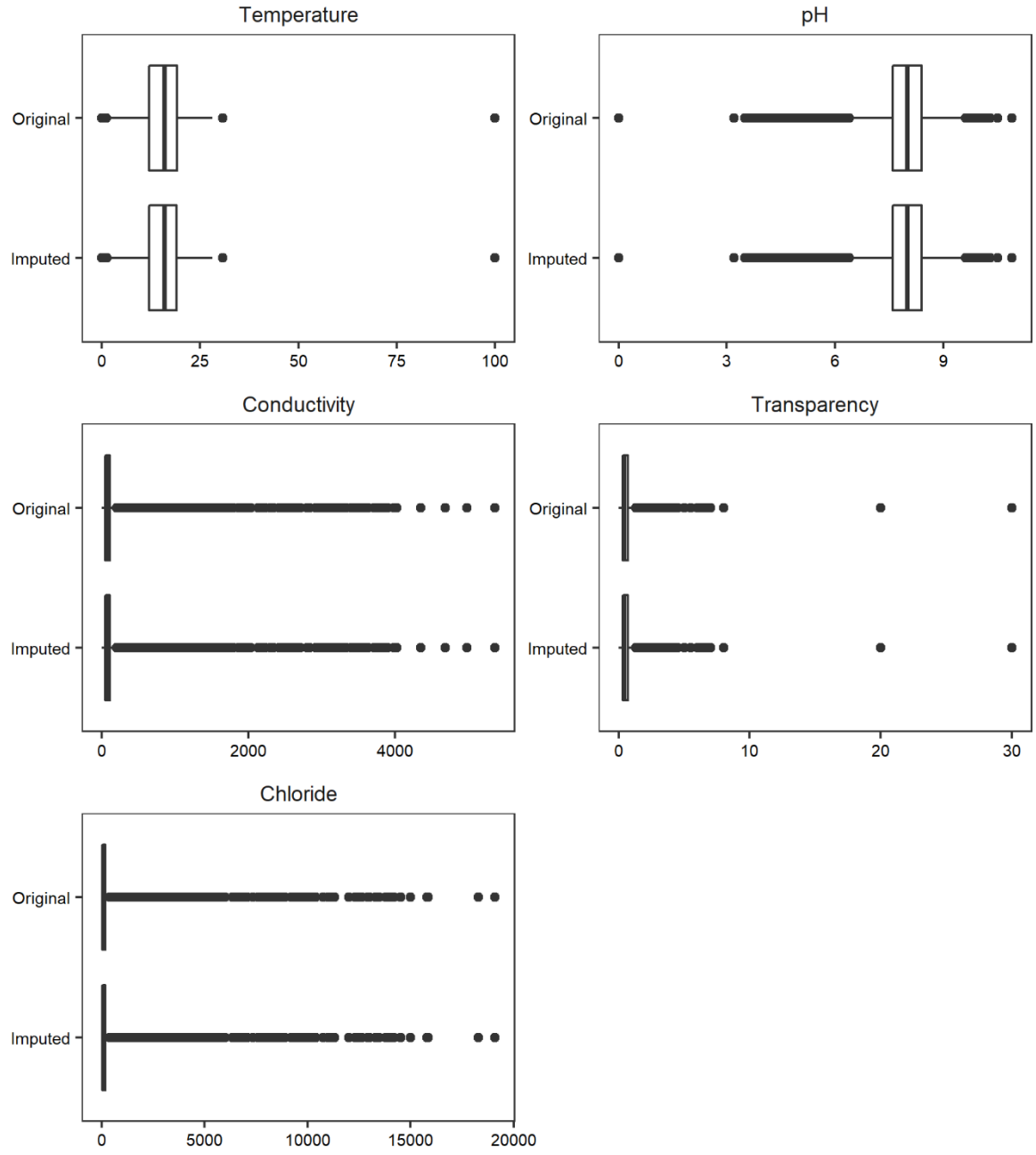

Figure B.16: Variable distributions before and after imputation by* missForest *(*mF*). *Replacement of missing values was performed for 269 data points in a data set with 5 variables and 5385 instances (hence, 1 % missing values).*

B.4.2 Case 2: Large data set with high degree of missing data

In Chapter 4, it was indicated that variable removal supported the decrease in missing data within the common data (see Figure A.4). However, it also showed a rapid decrease in both sample size and data dimensionality when less than 50 % missing data was aimed for. Therefore, this case considers the optimal data set, i.e. containing 10 variables and 17 264 instances (i.e. combination 1, Table 4.1), with 50 % missing data. Hence, in total 86 320 data points were artificially removed prior to imputation assessment. The variables within the data were chlorophyll *a* (µg·L^{-1}), chloride (mg·L^{-1}), conductivity (mS·m^{-1}), NH_4^+-N (mg·L^{-1}), oxygen (mg·L^{-1}), pH (-), PO_4^{3-}-P (mg·L^{-1}), temperature (°C), total phosphorus (mg·L^{-1}) and transparency (m).

B.4.2.1 Imputed values

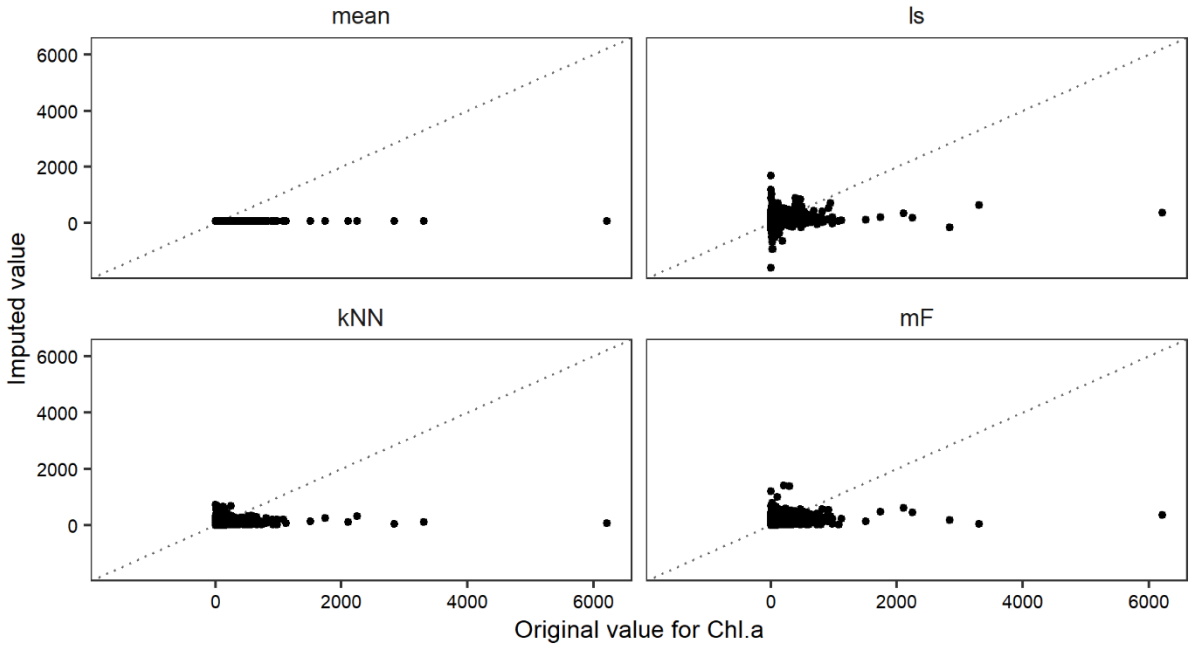

Figure B.17: Imputation of chlorophyll a by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are µg·L^{-1}.*

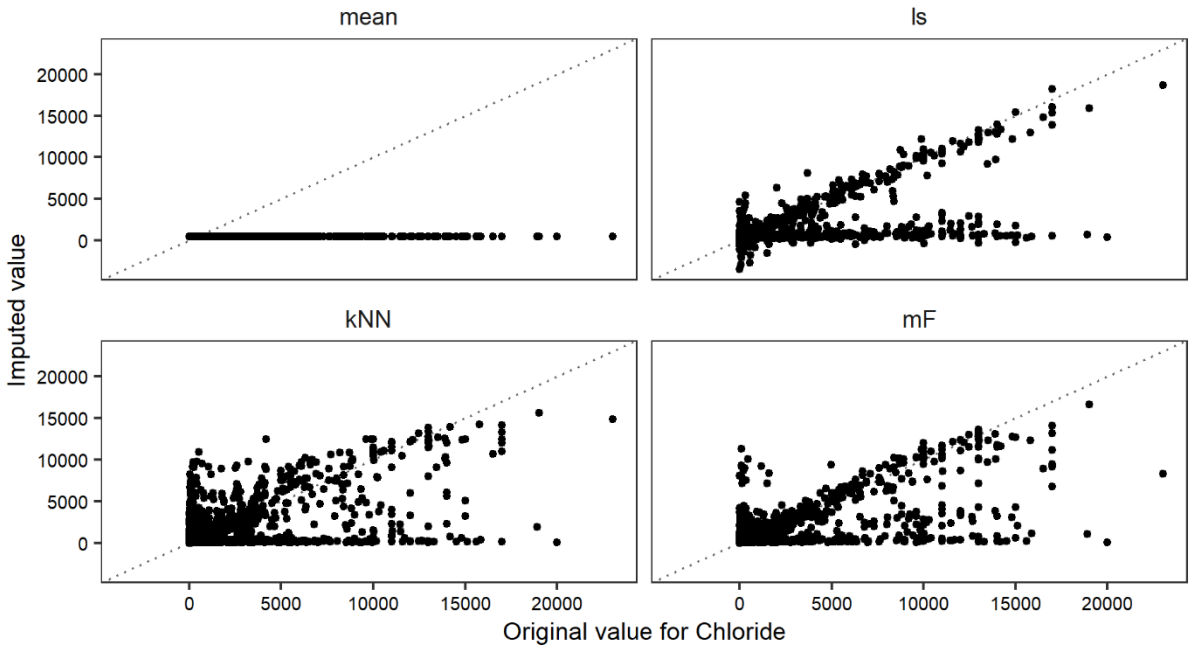

Figure B.18: Imputation of chloride by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mg·L⁻¹.*

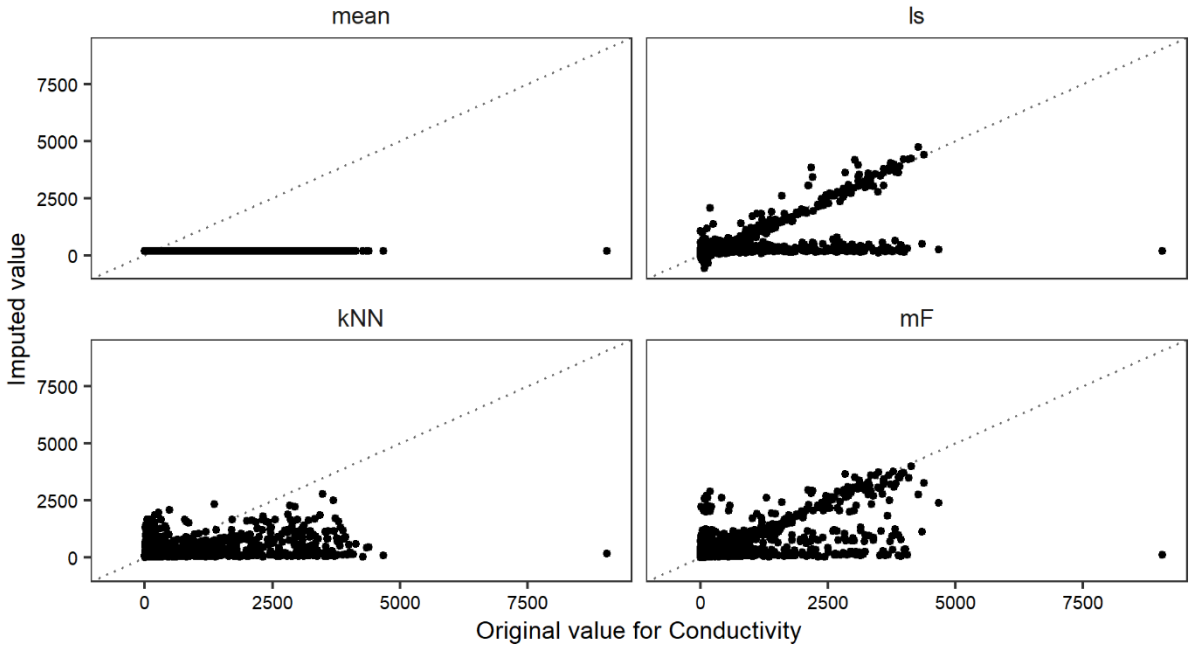

Figure B.19: Imputation of conductivity by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mS·m⁻¹.*

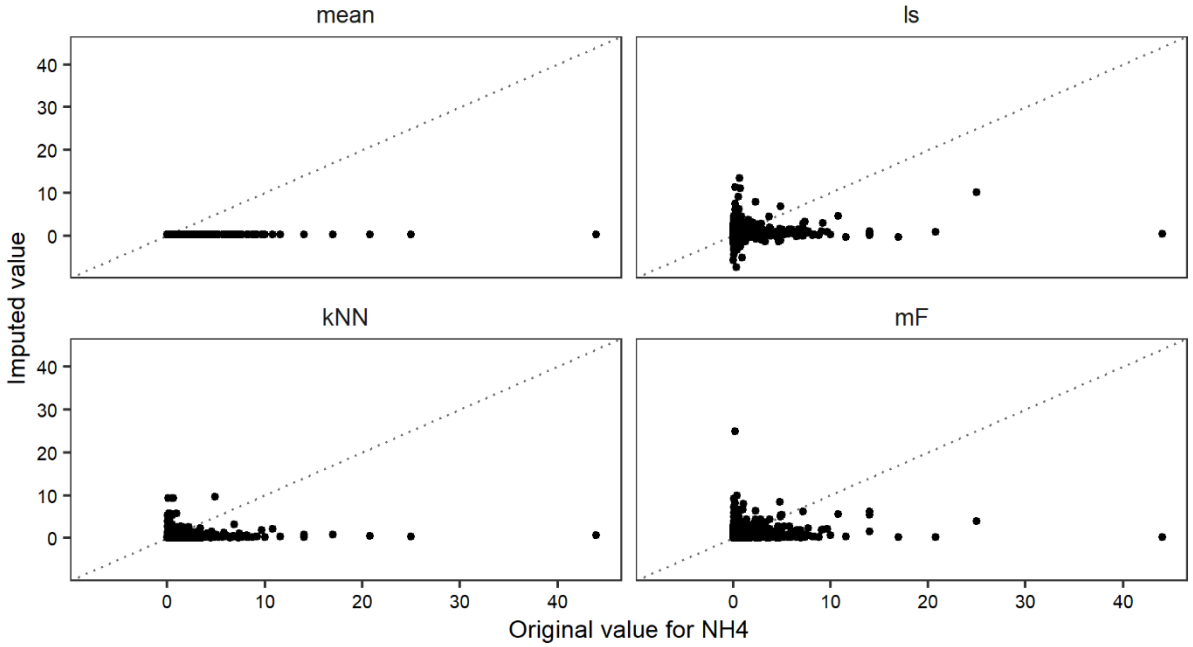

Figure B.20: Imputation of ammonium-nitrogen (NH_4^+-N) by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mg·L⁻¹.*

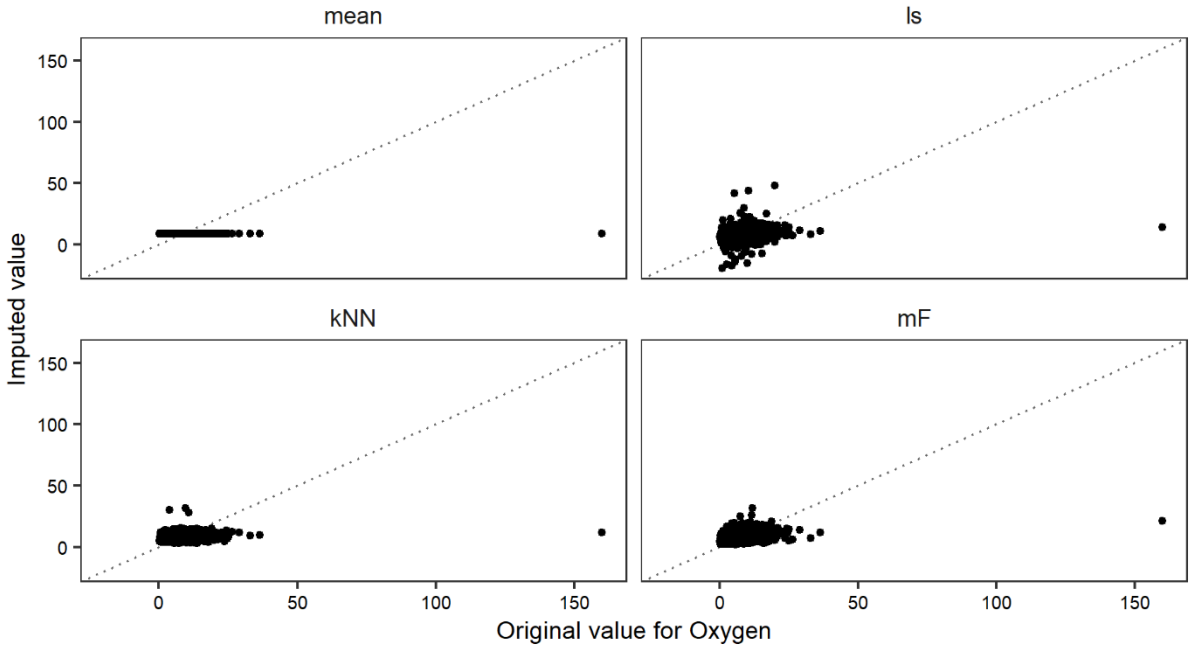

Figure B.21: Imputation of oxygen by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mg·L⁻¹.*

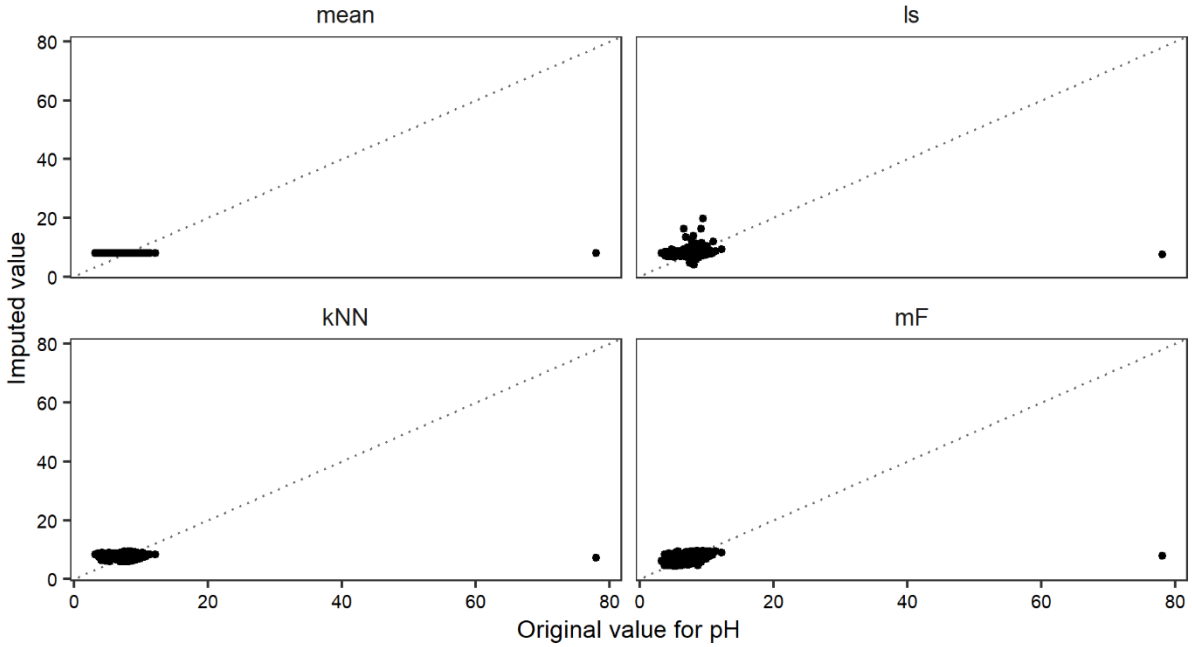

Figure B.22: Imputation of pH by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values).*

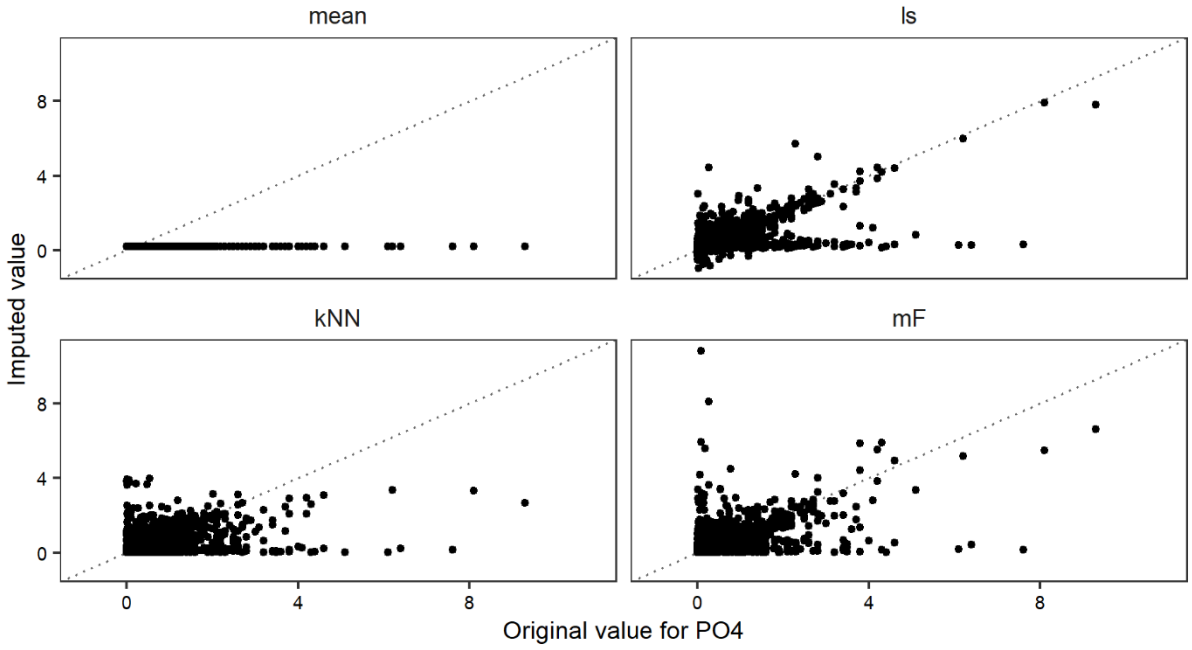

Figure B.23: Imputation of phosphate-phosphorus (PO_4^{3-}-P) by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mg·L⁻¹.*

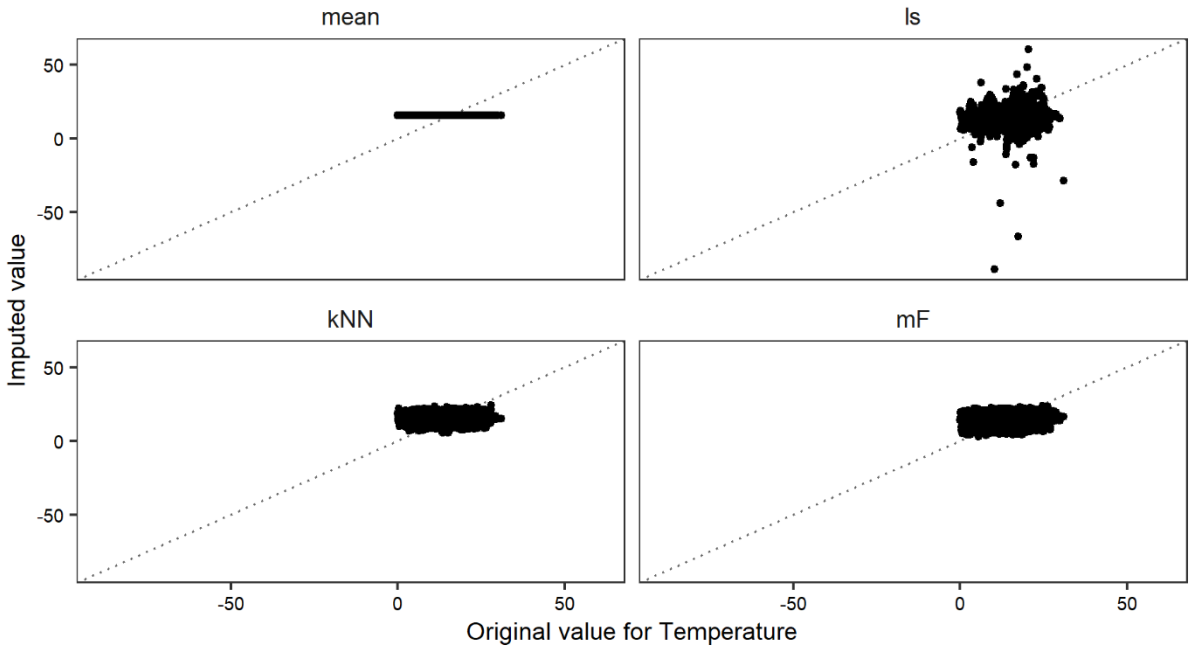

Figure B.24: Imputation of temperature by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are °C.*

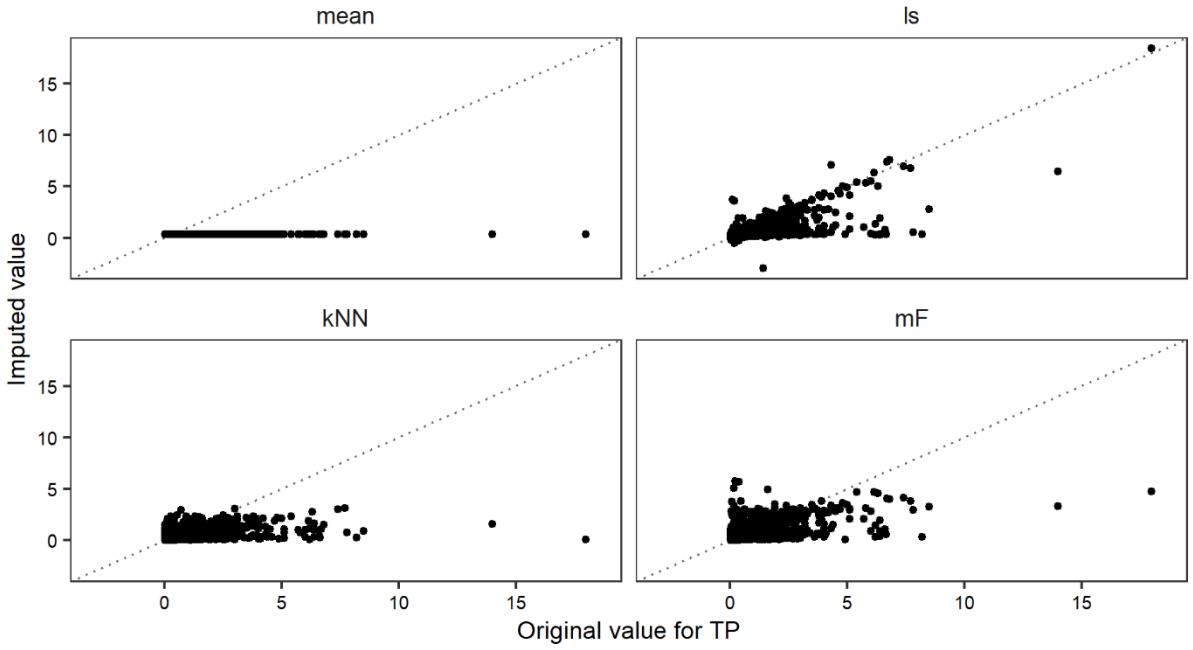

Figure B.25: Imputation of total phosphorus by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are mg·L⁻¹.*

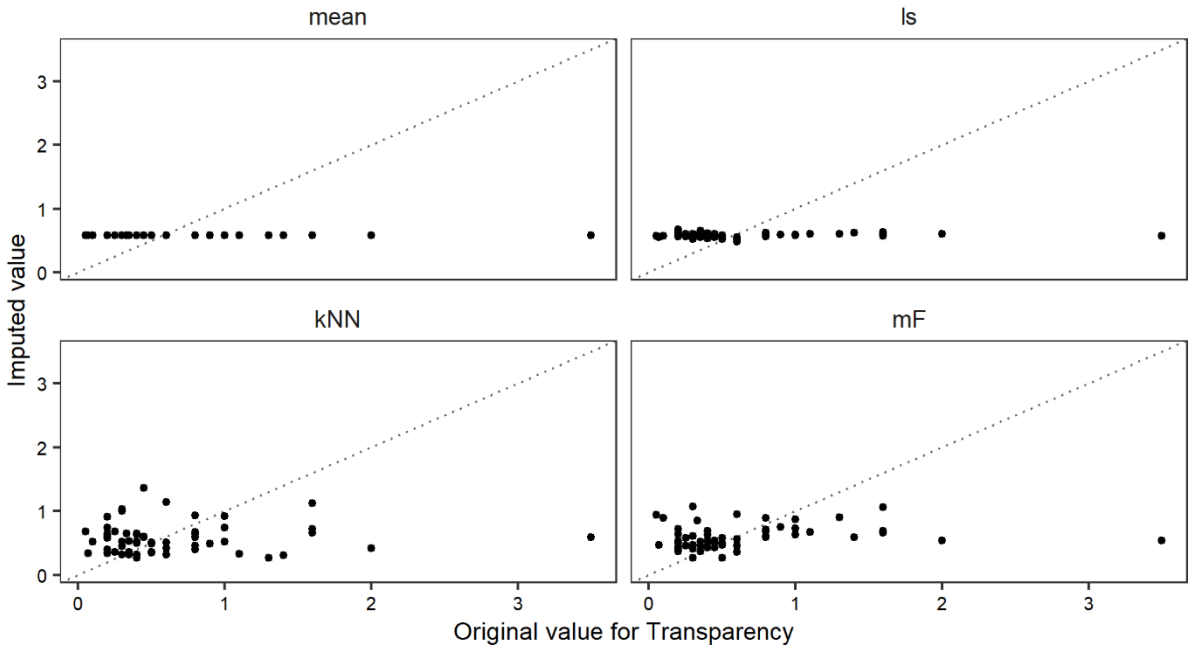

Figure B.26: Imputation of transparency by four imputation techniques. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values). Units are m.*

B.4.2.2 Variable distributions

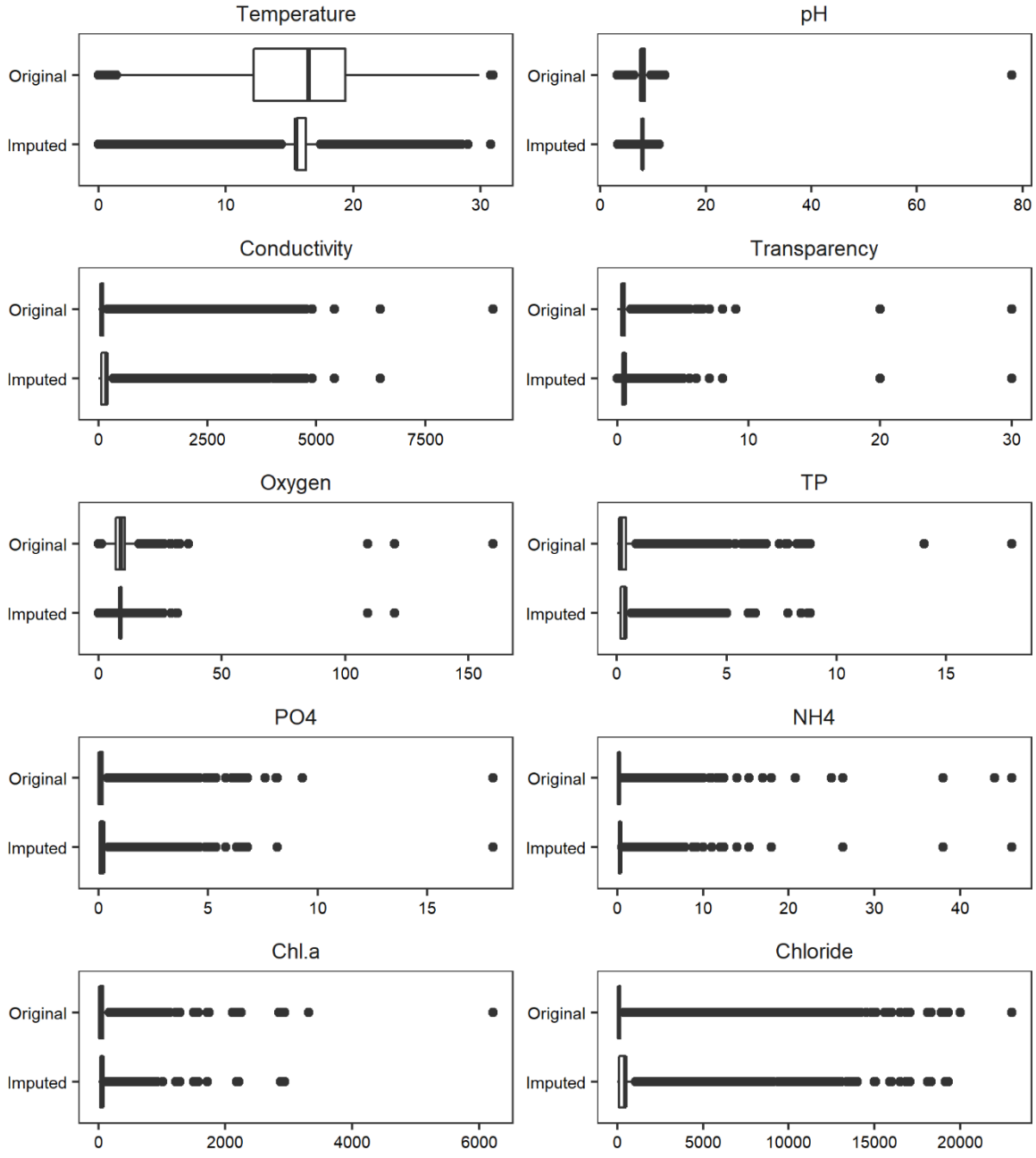

Figure B.27: Variable distributions before and after imputation by the mean. *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values).*

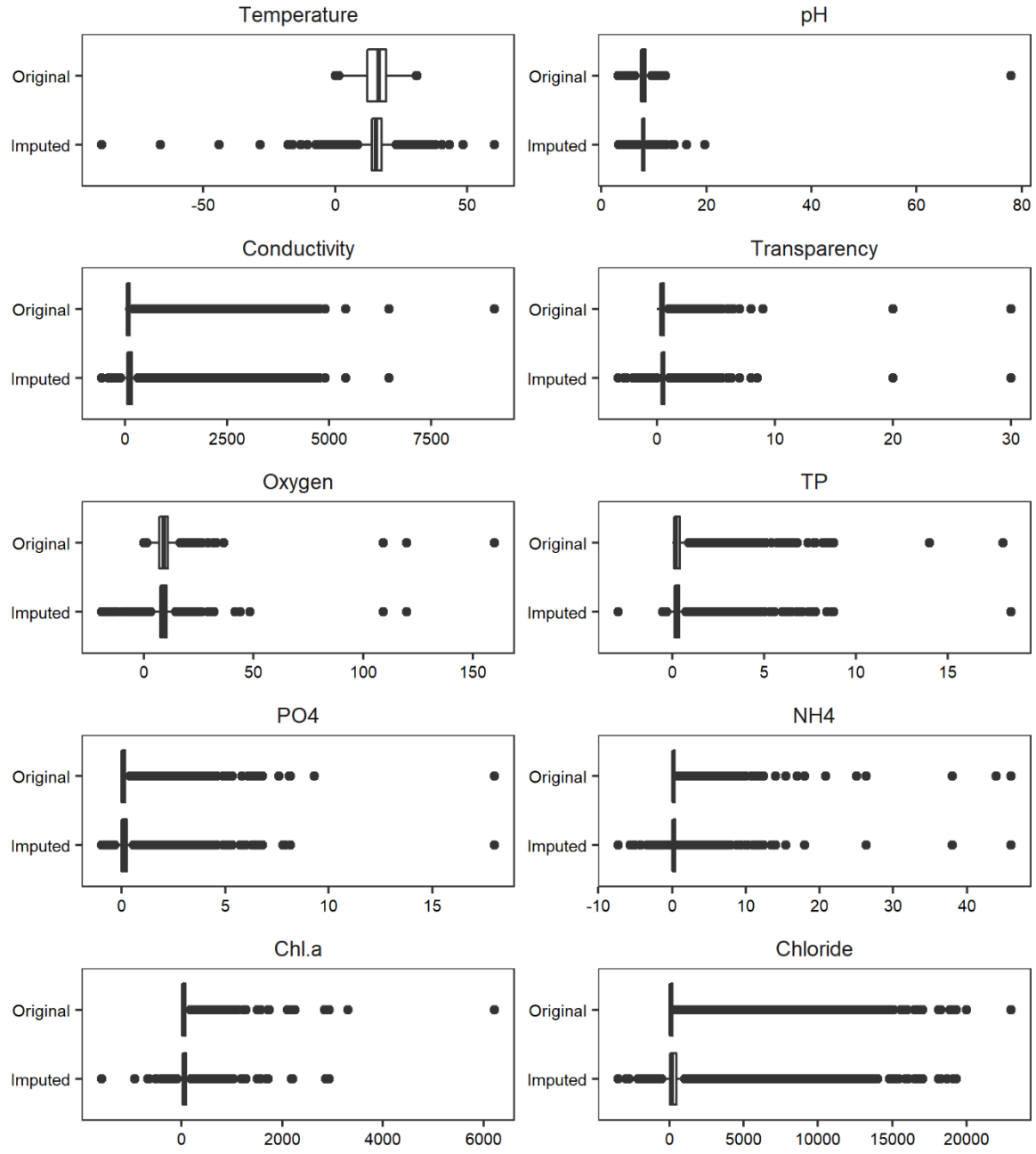

Figure B.28: Variable distributions before and after imputation by least squares regression (ls). *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values).*

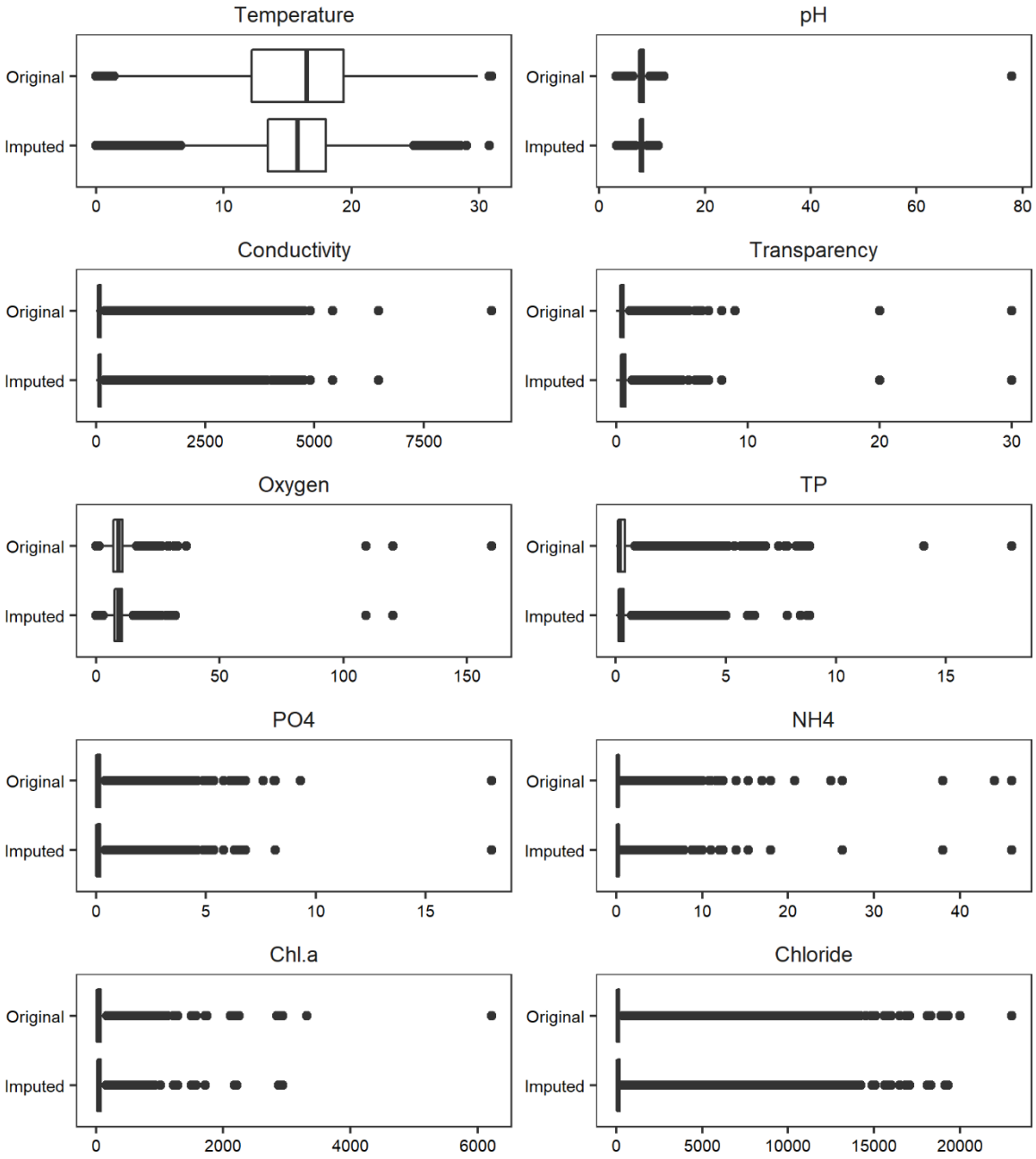

***Figure B.29: Variable distributions before and after imputation by* k *nearest neighbours* (kNN).** *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values).*

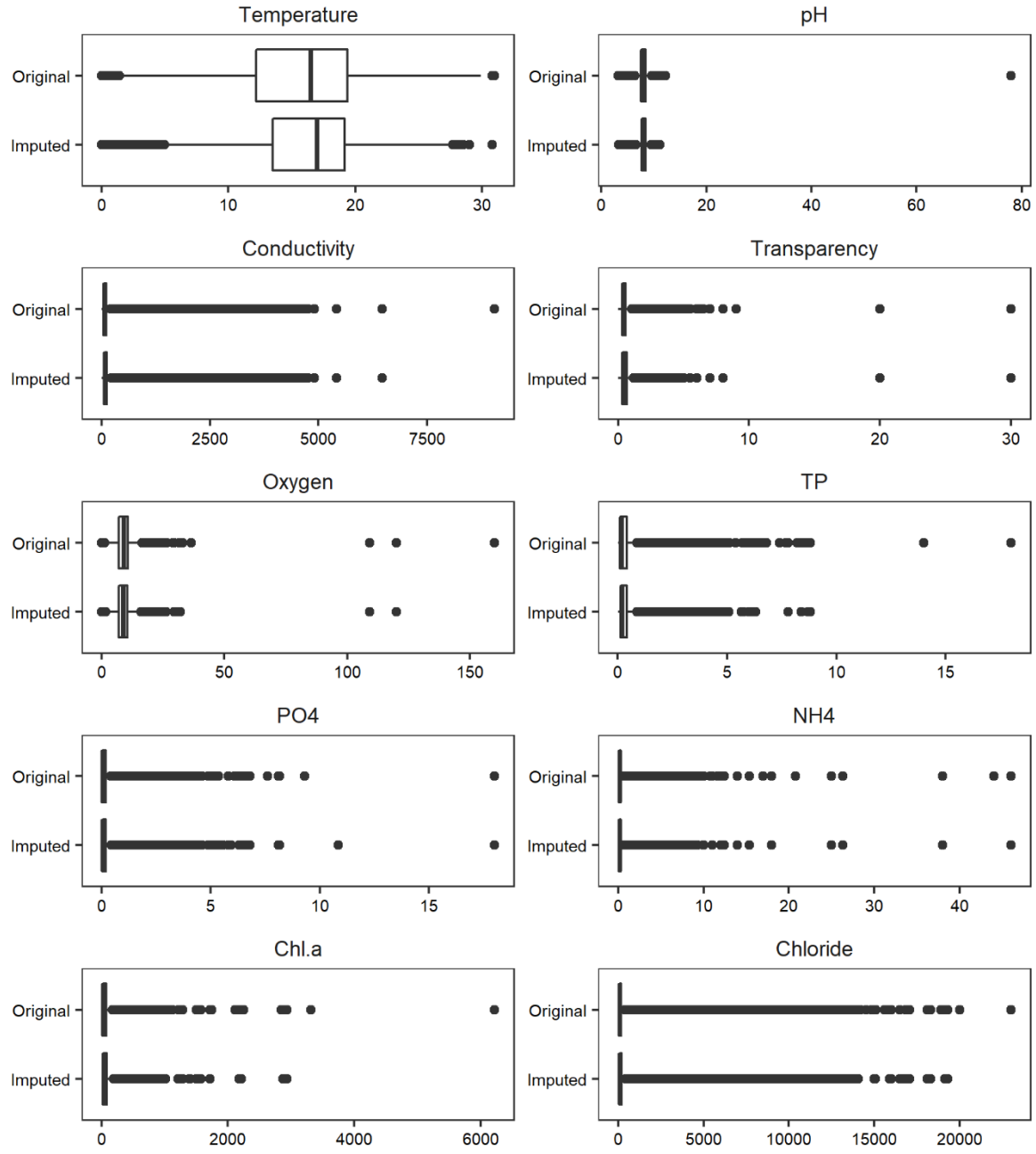

Figure B.30: Variable distributions before and after imputation by missForest (mF). *Replacement of missing values was performed for 86 320 data points in a data set with 10 variables and 17 264 instances (hence, 50 % missing values).*

B.4.3 Overall observations from the case studies

Imputation accuracy of the data by the four different techniques showed to differ between the percentage missing data (MD), the techniques and variables (see Figure B.8 up to Figure B.30). More specifically, imputed values obtained via *mean* imputation provided distinctly different patterns compared to *ls*, *kNN* and *mF*. However, this discrepancy seemed to be range-dependent as imputation patterns for variables with an extensive range clearly differed from the patterns obtained for variables with a confined range. For instance, chloride ranges from 0 up to 20 000 $mg \cdot L^{-1}$ and showed to be accurately imputed by *ls*, *kNN* and *mF* at 1 % MD (Figure B.8), while a lower accuracy could be inferred at 50 % MD (Figure B.18). Yet, clear differences were still present with *mean*, while *kNN* showed to be more prone to overestimate missing values compared to *ls* and *mF*. Similar observations can be done for conductivity at 1 % (Figure B.9) and 50 % (Figure B.19) MD. In contrast, imputations for pH at 1 % (Figure B.10) were generally better for *kNN* and *mF* compared to *ls*, which might be linked with the limited range (i.e. 6 up to 10). Due to an extreme value for pH, no clear statement could be made for 50 % MD (Figure B.22). These observations suggest that for variables with a limited range, *ls* provides relatively similar imputations compared to the mean (e.g. Figure B.11, Figure B.12 and Figure B.26), while *kNN* and *mF* provide relatively similar scores (e.g. Figure B.12, Figure B.24 and Figure B.25). However, no pairwise comparisons between the techniques has been performed.

The differences in imputed values explains the observed discrepancy at NRMSE level between *mean* and the remaining three techniques (see Figure 5.1 and Figure 5.4), with high underestimations by *mean* causing elevated NRMSE scores. More importantly, the imputation of the mean can be clearly observed in Figure B.27, which illustrates a clear narrowing effect on temperature values within the data. This narrowing is also observed for *ls* (Figure B.28), *kNN* (Figure B.29) and *mF* (Figure B.30), though in a less distinct manner. In contrast, narrowing could not be observed with only 1 % MD (see Figure B.13, Figure B.14, Figure B.15 and Figure B.16). Due to the extent of most variables, no clear effects originating from data imputation could be distinguished.

The relatively similar patterns obtained for *ls*, *kNN* and *mF* (see Figure B.8 up to Figure B.30) confirmed the high overlap in NRMSE observed in Figure 5.4. Still, minor differences could be observed due to the applied approach for imputing a missing value. For instance, *ls* is based on a global approach (i.e. considers all available data (Bø *et al.*, 2004)), causing relatively high bias and low variance. In contrast, *kNN* and *mF* are local approaches and rely on fractions of the original data to impute the missing value, thereby causing less bias and higher variance. Unfortunately, these observations do not allow to create additional performance distinction between the considered techniques, although improvement is expected to be limited for most data sets due to the relative overlap among the obtained NRMSE scores.

B.5 Linear Mixed Effects Models

B.5.1 Overall performance

Overall performance assessment considered the link between obtained imputation error and a full interaction model as specified in Equation B.1. For each factor, a coefficient was determined along with its deviation, confidence interval and contribution significance, as summarised in Table B.4.

$$NRMSE = \beta_0 + \beta_{ls} \cdot Method_{ls} + \beta_{kNN} \cdot Method_{kNN} + \beta_{mF} \cdot Method_{mF} + \beta_{MD} \cdot MD +$$
$$\beta_{Inst} \cdot N_{inst} + \beta_{Var} \cdot N_{var} +$$
$$\beta_{ls:MD} \cdot Method_{ls}{:}\, MD + \beta_{kNN:MD} \cdot Method_{kNN}{:}\, MD + \beta_{mF:MD} \cdot Method_{mF}{:}\, MD +$$
$$\beta_{ls:Inst} \cdot Method_{ls}{:}\, N_{inst} + \beta_{kNN:Inst} \cdot Method_{kNN}{:}\, N_{inst} + \beta_{mF:Inst} \cdot Method_{mF}{:}\, N_{inst} +$$
$$\beta_{ls:Var} \cdot Method_{ls}{:}\, N_{var} + \beta_{kNN:Var} \cdot Method_{kNN}{:}\, N_{var} + \beta_{mF:Var} \cdot Method_{mF}{:}\, N_{var} +$$
$$\beta_{MD:Inst} \cdot MD{:}\, N_{inst} + \beta_{MD:Var} \cdot MD{:}\, N_{var} + \beta_{Inst:Var} \cdot N_{inst}{:}\, N_{var} +$$
$$\beta_{ls:MD:Inst} \cdot Method_{ls}{:}\, MD{:}\, N_{inst} + \beta_{kNN:MD:Inst} \cdot Method_{kNN}{:}\, MD{:}\, N_{inst} +$$
$$\beta_{mF:MD:Inst} \cdot Method_{mF}{:}\, MD{:}\, N_{inst} + \beta_{ls:MD:Var} \cdot Method_{ls}{:}\, MD{:}\, N_{var} +$$
$$\beta_{kNN:MD:Inst} \cdot Method_{kNN}{:}\, MD{:}\, N_{var} + \beta_{mF:MD:Inst} \cdot Method_{mF}{:}\, MD{:}\, N_{var} +$$
$$\beta_{ls:MD:Inst} \cdot Method_{ls}{:}\, MD{:}\, N_{inst} + \beta_{ls:Inst:Var} \cdot Method_{ls}{:}\, N_{inst}{:}\, N_{var} +$$
$$\beta_{kNN:Inst:Var} \cdot Method_{kNN}{:}\, N_{inst}{:}\, N_{var} + \beta_{mF:Inst:Var} \cdot Method_{mF}{:}\, N_{inst}{:}\, N_{var} +$$
$$\beta_{MD:Inst:Var} \cdot MD{:}\, N_{inst}{:}\, N_{var} + \beta_{ls:MD:Inst:Var} \cdot Method_{ls}{:}\, MD{:}\, N_{inst}{:}\, N_{var} +$$
$$\beta_{kNN:MD:Inst:Var} \cdot Method_{kNN}{:}\, MD{:}\, N_{inst}{:}\, N_{var} +$$
$$\beta_{mF:MD:Inst:Var} \cdot Method_{mF}{:}\, MD{:}\, N_{inst}{:}\, N_{var} \qquad \text{(Equation B.1)}$$

Table B.4: Summary of coefficients within the overall mixed effect model, linking performance (NRMSE) with imputation method, fraction missing data, fraction of instances and number of variables (NRMSE~Method*MD*Inst*Var + (1|n)). *Each coefficient is supplemented with its standard deviation (SD), 95 % confidence interval (CI2.5% - CI97.5%) and contribution significance. Codes: ls: least squares; kNN: k nearest neighbours, mF: missForest algorithm; MD: fraction missing data; Inst: fraction of instances; Var: fraction of variables.*

Effect	Coefficient	SD	CI2.5%	CI97.5%	*p*
Intercept	0.903	0.049	0.808	0.997	<0.001
ls	-0.620	0.049	-0.716	-0.524	<0.001
kNN	-0.746	0.049	-0.842	-0.650	<0.001
mF	-0.712	0.049	-0.808	-0.616	<0.001
MD	0.169	0.130	-0.080	0.419	0.186
Inst	0.115	0.071	-0.023	0.254	0.104
Var	0.113	0.045	0.026	0.201	0.012
ls:MD	0.682	0.129	0.430	0.934	<0.001
kNN:MD	1.672	0.129	1.419	1.924	<0.001
mF:MD	0.941	0.129	0.689	1.193	<0.001
ls:Inst	-0.072	0.072	-0.212	0.068	0.317
kNN:Inst	0.042	0.072	-0.098	0.181	0.563
mF:Inst	0.033	0.072	-0.106	0.174	0.637
MD:Inst	-0.193	0.187	-0.557	0.172	0.302
ls:Var	-0.122	0.045	-0.211	-0.033	0.007
kNN:Var	0.022	0.045	-0.067	0.110	0.633
mF:Var	-0.034	0.045	-0.122	0.055	0.460
MD:Var	-0.272	0.119	-0.503	-0.041	0.022
Inst:Var	-0.183	0.066	-0.312	-0.055	0.005
ls:MD:Inst	0.198	0.189	-0.170	0.567	0.294
kNN:MD:Inst	-0.701	0.189	-1.069	-0.332	<0.001
mF:MD:Inst	0.042	0.189	-0.327	0.410	0.825
ls:MD:Var	0.191	0.120	-0.043	0.424	0.112
kNN:MD:Var	-0.506	0.120	-0.739	-0.272	<0.001
mF:MD:Var	-0.056	0.120	-0.289	0.178	0.642
ls:Inst:Var	0.101	0.066	-0.029	0.230	0.129
kNN:Inst:Var	-0.017	0.066	-0.147	0.112	0.794
mF:Inst:Var	-0.025	0.066	-0.155	0.104	0.705
MD:Inst:Var	0.309	0.173	-0.029	0.646	0.075
ls:MD:Inst:Var	-0.258	0.175	-0.599	0.083	0.141
kNN:MD:Inst:Var	0.466	0.175	0.125	0.807	0.008
mF:MD:Inst:Var	-0.092	0.175	-0.433	0.249	0.597

B.5.2 Baseline performance

Baseline performance assessment considered the link between obtained imputation error related to $\boldsymbol{D_{opt}}$ and a full interaction model as specified in Equation B.2. For each factor, a coefficient was determined along with its deviation, confidence interval and contribution significance, as summarised in Table B.5.

$$NRMSE = \beta_0 + \beta_{ls} \cdot Method_{ls} + \beta_{kNN} \cdot Method_{kNN} + \beta_{mF} \cdot Method_{mF} + \beta_{MD} \cdot MD + \beta_{ls:MD} \cdot Method_{ls}: MD + \beta_{kNN:MD} \cdot Method_{kNN}: MD + \beta_{mF:MD} \cdot Method_{mF}: MD \quad \text{(Equation B.2)}$$

Table B.5: Summary of coefficients within the baseline mixed effect model, linking performance (NRMSE) with imputation method and fraction missing data (NRMSE~Method*MD + (1\|n)). *Each coefficient is supplemented with its standard deviation (SD), 95 % confidence interval (CI2.5% - CI97.5%) and contribution significance. Codes: ls: least squares; kNN: k nearest neighbours, mF: missForest algorithm; MD: fraction missing data.*

Effect	Coefficient	SD	CI2.5%	CI97.5%	*p*
Intercept	0.966	0.011	0.946	0.987	<0.001
ls	-0.710	0.010	-0.730	-0.689	<0.001
kNN	-0.716	0.010	-0.736	-0.696	<0.001
mF	-0.739	0.010	-0.759	-0.719	<0.001
MD	-0.0005	0.028	-0.055	0.054	0.986
ls:MD	0.822	0.027	0.769	0.874	<0.001
kNN:MD	1.080	0.027	1.027	1.132	<0.001
mF:MD	0.859	0.027	0.807	0.912	<0.001

B.5.3 Sample size variability

Sample size variability performance assessment considered the link between obtained imputation error related to $\boldsymbol{D_{opt}}$ and a full interaction model as specified in Equation B.3. For each factor, a coefficient was determined along with its deviation, confidence interval and contribution significance, as summarised in Table B.6.

$$NRMSE = \beta_0 + \beta_{ls} \cdot Method_{ls} + \beta_{kNN} \cdot Method_{kNN} + \beta_{mF} \cdot Method_{mF} + \beta_{MD} \cdot MD + \beta_{Inst} \cdot N_{inst} + \beta_{ls:MD} \cdot Method_{ls}: MD + \beta_{kNN:MD} \cdot Method_{kNN}: MD + \beta_{mF:MD} \cdot Method_{mF}: MD + \beta_{ls:Inst} \cdot Method_{ls}: N_{inst} + \beta_{kNN:Inst} \cdot Method_{kNN}: N_{inst} + \beta_{mF:Inst} \cdot Method_{mF}: N_{inst} + \beta_{MD:Inst} \cdot MD: N_{inst} + \beta_{ls:MD:Inst} \cdot Method_{ls}: MD: N_{inst} + \beta_{kNN:MD:Inst} \cdot Method_{kNN}: MD: N_{inst} + \beta_{mF:MD:Inst} \cdot Method_{mF}: MD: N_{inst} \quad \text{(Equation B.3)}$$

Table B.6: Summary of coefficients within the mixed effect model for sample size variability, linking performance (NRMSE) with imputation method, fraction missing data and fraction of instances (NRMSE~Method*MD*Inst + (1|n)). *Each coefficient is supplemented with its standard deviation (SD), 95 % confidence interval (CI2.5% - CI97.5%) and contribution significance. Codes: ls: least squares; kNN: k nearest neighbours, mF: missForest algorithm; MD: fraction missing data; Inst: fraction of instances.*

Effect	Coefficient	SD	CI2.5%	CI97.5%	*p*
Intercept	0.985	0.026	0.934	1.036	<0.001
ls	-0.634	0.029	-0.690	-0.577	<0.001
kNN	-0.622	0.029	-0.678	-0.565	<0.001
mF	-0.749	0.029	-0.806	-0.693	<0.001
MD	-0.032	0.069	-0.166	0.102	0.637
Inst	-0.022	0.038	-0.096	0.053	0.569
ls:MD	0.716	0.076	0.567	0.865	<0.001
kNN:MD	0.937	0.076	0.789	1.086	<0.001
mF:MD	0.915	0.076	0.767	1.064	<0.001
ls:Inst	-0.091	0.042	-0.174	-0.009	0.032
kNN:Inst	-0.096	0.042	-0.179	-0.014	0.023
mF:Inst	0.013	0.042	-0.070	0.095	0.760
MD:Inst	0.038	0.101	-0.158	0.234	0.704
ls:MD:Inst	0.124	0.112	-0.094	0.341	0.269
kNN:MD:Inst	0.134	0.112	-0.083	0.351	0.230
mF:MD:Inst	-0.073	0.112	-0.290	0.144	0.513

B.5.4 Dimensionality variability

Sample size variability performance assessment considered the link between obtained imputation error related to $\boldsymbol{D_{opt}}$ and a full interaction model as specified in Equation B.4. For each factor, a coefficient was determined along with its deviation and confidence interval, as summarised in Table B.7.

$$NRMSE = \beta_0 + \beta_{ls} \cdot Method_{ls} + \beta_{kNN} \cdot Method_{kNN} + \beta_{mF} \cdot Method_{mF} +$$
$$\beta_{MD} \cdot MD + \beta_{Var} \cdot N_{var} + \beta_{ls:MD} \cdot Method_{ls}: MD + \beta_{kNN:MD} \cdot Method_{kNN}: MD +$$
$$\beta_{mF:MD} \cdot Method_{mF}: MD + \beta_{ls:Var} \cdot Method_{ls}: N_{var} + \beta_{kNN:Var} \cdot Method_{kNN}: N_{var} +$$
$$\beta_{mF:Var} \cdot Method_{mF}: N_{var} + \beta_{MD:Inst} \cdot MD: N_{inst} + \beta_{ls:MD:Var} \cdot Method_{ls}: MD: N_{var} +$$
$$\beta_{kNN:MD:Var} \cdot Method_{kNN}: MD: N_{var} +$$
$$\beta_{mF:MD:Var} \cdot Method_{mF}: MD: N_{var} \qquad \text{(Equation B.4)}$$

Table B.7: Summary of coefficients within the overall mixed effect model, linking performance (NRMSE) with imputation method, fraction missing data and number of variables (NRMSE~Method*MD+Var+Method:Var + (1|n)). *Each coefficient is supplemented with its standard deviation (SD), 95 % confidence interval (CI2.5% - CI97.5%) and contribution significance. Codes: ls: least squares; kNN: k nearest neighbours, mF: missForest algorithm; MD: fraction missing data; Var: fraction of variables.*

Effect	Coefficient	SD	CI2.5%	CI97.5%	*p*
Intercept	1.001	0.018	0.971	1.041	<0.001
ls	-0.670	0.018	-0.704	-0.636	<0.001
kNN	-0.686	0.018	-0.720	-0.652	<0.001
mF	-0.668	0.018	-0.702	-0.634	<0.001
MD	-0.003	0.047	-0.096	0.089	0.942
Var	-0.050	0.017	-0.082	-0.018	0.003
ls:MD	0.838	0.046	0.748	0.928	<0.001
kNN:MD	1.023	0.046	0.933	1.113	<0.001
mF:MD	0.922	0.046	0.832	1.011	<0.001
ls:Var	-0.044	0.016	-0.076	-0.013	0.007
kNN:Var	-0.025	0.016	-0.057	0.006	0.123
mF:Var	-0.067	0.016	-0.098	-0.035	<0.001
MD:Var	0.0017	0.044	-0.083	0.087	0.968
ls:MD:Var	-0.018	0.043	-0.101	0.065	0.670
kNN:MD:Var	-0.039	0.043	-0.122	0.044	0.361
mF:MD:Var	-0.092	0.043	-0.175	-0.009	0.031

C

Supportive Information for Chapter 6 – Threshold selection for data pre-processing

C.1 Data reduction

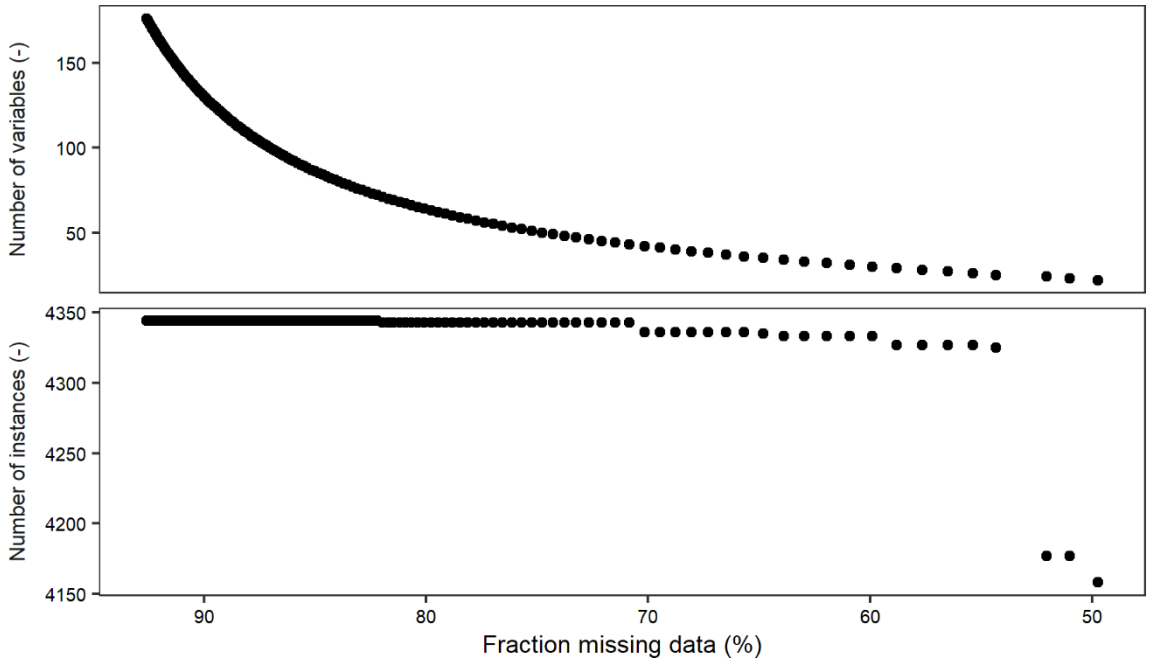

Figure C.1: Reduction of missing data by stepwise removal of variables or instances. *Variable removal often caused the highest drop in fraction missing data and is therefore more frequently applied than instance removal. Data reduction was performed until about 50 % of the data was available to estimate the remaining 50 % of the data.*

Table C.1: Overview of the variables remaining after data reduction and imputation. *Data reduction resulted in a total of 20 variables remaining (see also Figure C.1). These variables were subsequently considered for further data pre-processing, especially during selection of relevant explanatory variables.*

Variable	Unit	Variable	Unit
Temperature	°C	Phosphate-P	$mg \cdot L^{-1}$
pH	-	Kjeldahl nitrogen	$mg \cdot L^{-1}$
Conductivity	$mS \cdot m^{-1}$	Nitrite-N	$mg \cdot L^{-1}$
Oxygen saturation	%	Calcium	$mg \cdot L^{-1}$
Chloride	$mg \cdot L^{-1}$	Sulphate	$mg \cdot L^{-1}$
Oxygen	$mg \cdot L^{-1}$	BOD_5	$mg \cdot L^{-1}$
Transparency	m	Magnesium	$mg \cdot L^{-1}$
Ammonium-N	$mg \cdot L^{-1}$	Potassium	$mg \cdot L^{-1}$
Total phosphorus	$mg \cdot L^{-1}$	Sodium	$mg \cdot L^{-1}$
Nitrate-N	$mg \cdot L^{-1}$	Chlorophyll *a*	$\mu g \cdot L^{-1}$

Table C.2: Overview of the macrophytes considered in this study. *For each macrophyte, its prevalence, main growth form and native/alien background are provided. Not all macrophytes tend to occur in completely waterlogged systems, but were included to represent the wetland systems. Native or alien origin is considered with respect to western Europe.*

Macrophyte	Prevalence (%)	Growth form	Origin
Acorus calamus	3.94	Emergent	Alien
Alopecurus geniculatus	2.43	Emergent	Native
Berula erecta	5.19	Emergent	Native
Bidens tripartita	2.45	Emergent	Native
Butomus umbellatus	6.18	Emergent	Native
Callitriche platycarpa	5.32	Submerged	Native
Carex acuta	2.41	Emergent	Native
Carex pseudocyperus	2.62	Emergent	Native
Carex riparia	4.11	Emergent	Native
Ceratophyllum demersum	18.47	Submerged	Native
Eleocharis palustris	4.67	Emergent	Native
Elodea nuttallii	21.14	Submerged	Alien
Equisetum palustre	2.69	Emergent	Native
Eupatorium cannabinum	4.86	Emergent	Native
Filipendula ulmaria	3.13	Emergent	Native
Galium aparine	2.55	Emergent	Native
Glyceria fluitans	7.77	Emergent	Alien
Glyceria maxima	28.55	Emergent	Native
Iris pseudacorus	18.11	Emergent	Native
Juncus articulatus	4.26	Emergent	Native
Juncus effusus	12.29	Emergent	Native
Juncus inflexus	2.67	Emergent	Native
Lemna gibba	11.28	Floating	Native
Lemna minor	26.72	Floating	Native
Lemna minuta	3.39	Floating	Alien
Lemna trisulca	9.98	Submerged	Native
Lycopus europaeus	13.16	Emergent	Native
Lythrum salicaria	6.49	Emergent	Native
Mentha aquatica	10.51	Emergent	Native
Myosotis laxa	2.65	Emergent	Native
Myosotis scorpioides	8.71	Emergent	Native
Myriophyllum spicatum	4.23	Submerged	Native
Nasturtium microphyllum	3.10	Emergent	Native

(Continues on next page)

(Continued)

Macrophyte	Prevalence (%)	Growth form	Origin
Nuphar lutea	10.46	Floating	Native
Nymphaea alba	6.25	Floating	Native
Nymphoides peltata	2.96	Floating	Native
Persicaria amphibia	11.90	Floating	Native
Phalaris arundinacea	11.47	Emergent	Native
Phragmites australis	41.34	Emergent	Native
Potamogeton crispus	2.45	Submerged	Native
Potamogeton natans	3.32	Floating	Native
Potamogeton pectinatus	8.87	Submerged	Native
Potamogeton pusillus	5.22	Submerged	Native
Ranunculus circinatus	2.48	Submerged	Native
Ranunculus repens	5.22	Emergent	Native
Ranunculus sceleratus	5.80	Emergent	Native
Rorippa amphibia	6.57	Emergent	Native
Rumex hydrolapathum	11.28	Emergent	Native
Sagittaria sagittifolia	6.08	Emergent	Native
Sparganium emersum	3.44	Emergent	Native
Sparganium erectum	13.66	Emergent	Native
Sphagnum majus	29.00	Emergent	Native
Sphagnum pulchrum	12.48	Emergent	Native
Spirodela polyrhiza	18.69	Floating	Native
Stachys palustris	6.71	Emergent	Native
Symphytum officinale	4.79	Emergent	Native
Typha angustifolia	6.52	Emergent	Native
Typha latifolia	11.59	Emergent	Native

C.2 Effects of threshold selection

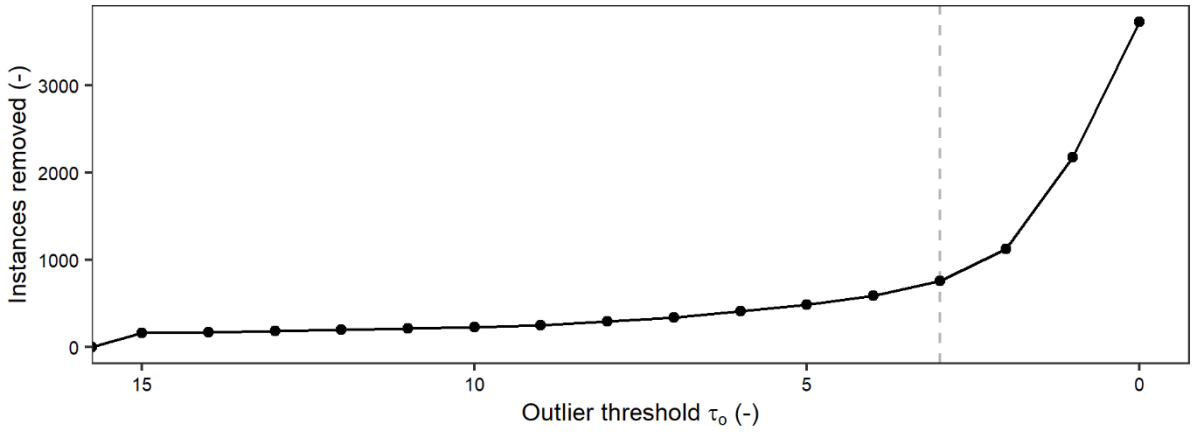

Figure C.2: Relation between the used threshold and number of instances removed. *With decreasing threshold values, more instances are considered as outlier and consequently removed from the data set. At first, the increase is relatively small, though becomes exponential when dropping below $\tau_o = 5$. Threshold selection of $\tau_o = 3$ (dashed grey line) causes the removal of 760 instances.*

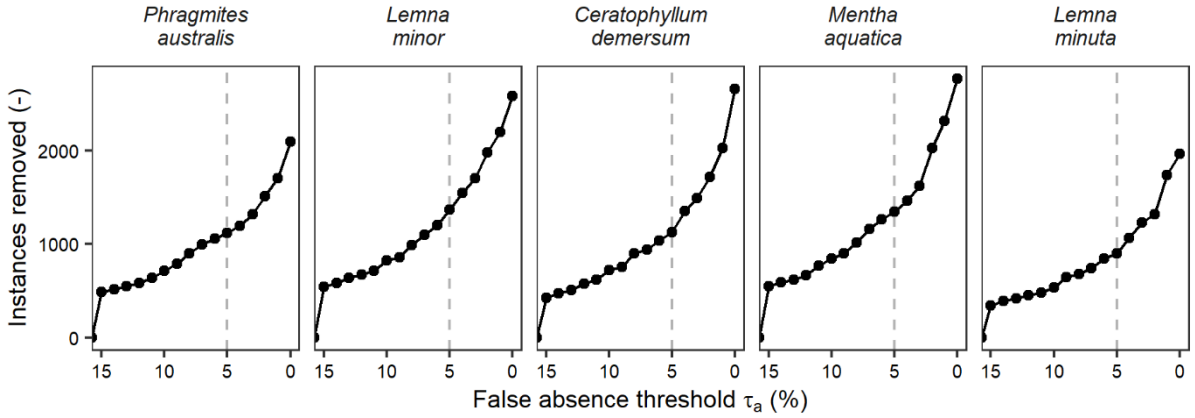

Figure C.3: Relation between the false absence threshold and number of instances removed. *With decreasing threshold values, exponentially more instances are considered as potential false absences. Implementation of a conservative threshold ($\tau_a = 15$ %) causes a relatively high number of instances to be removed, while selection of $\tau_a = 5$ % (dashed grey line) impedes the removal of too many instances.*

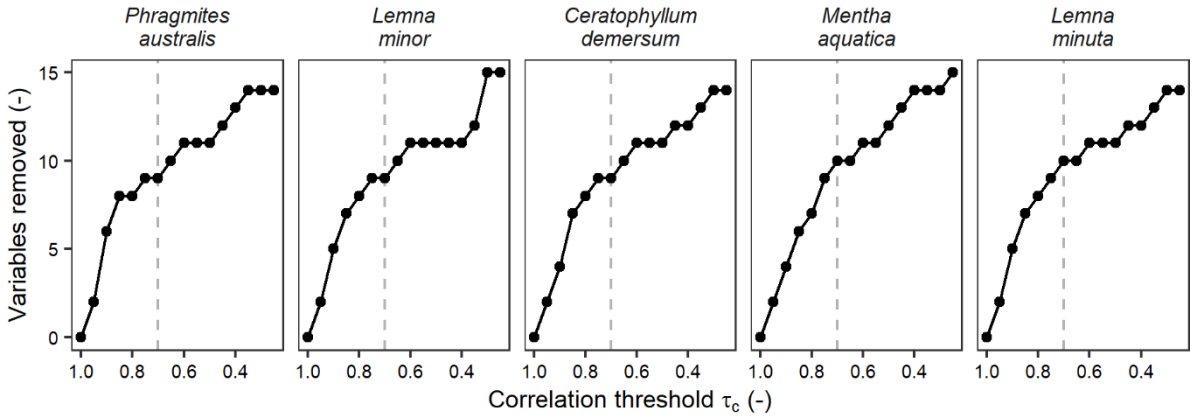

Figure C.4: Relation between the correlation threshold and the number of variables removed. *With decreasing threshold values, more variables are considered as being correlated. Even with conservative threshold scores (e.g. τ_c = 0.9), 5 or more variables are already being removed. Threshold selection at τ_c = 0.7 (dashed grey line) limits variable removal to only 10 variables being removed.*

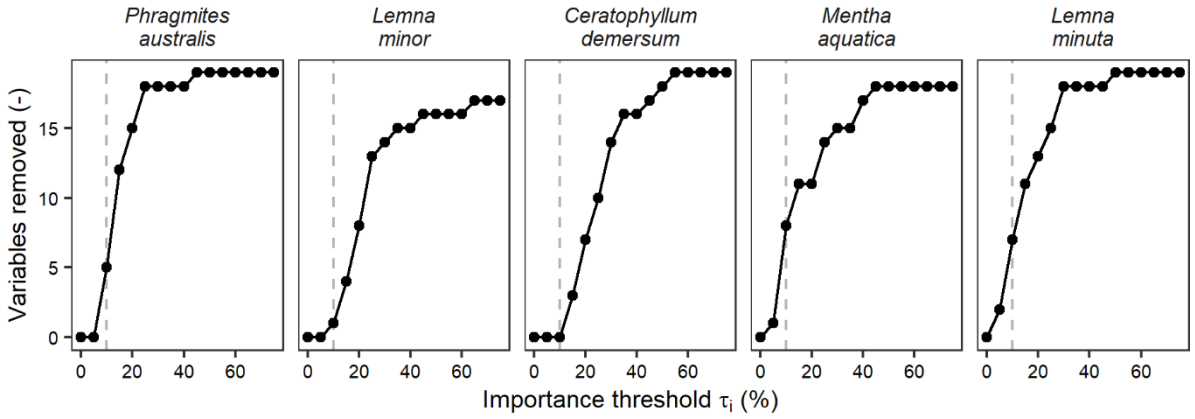

Figure C.5: Relation between the variable importance threshold and the number of variables removed. *With increasing threshold values, more variables are being considered as irrelevant. Even with conservative threshold scores (e.g. τ_i = 20 %), high numbers of variables are removed. Threshold selection at τ_i = 10 % (dashed grey line) limits overly excessive variable removal.*

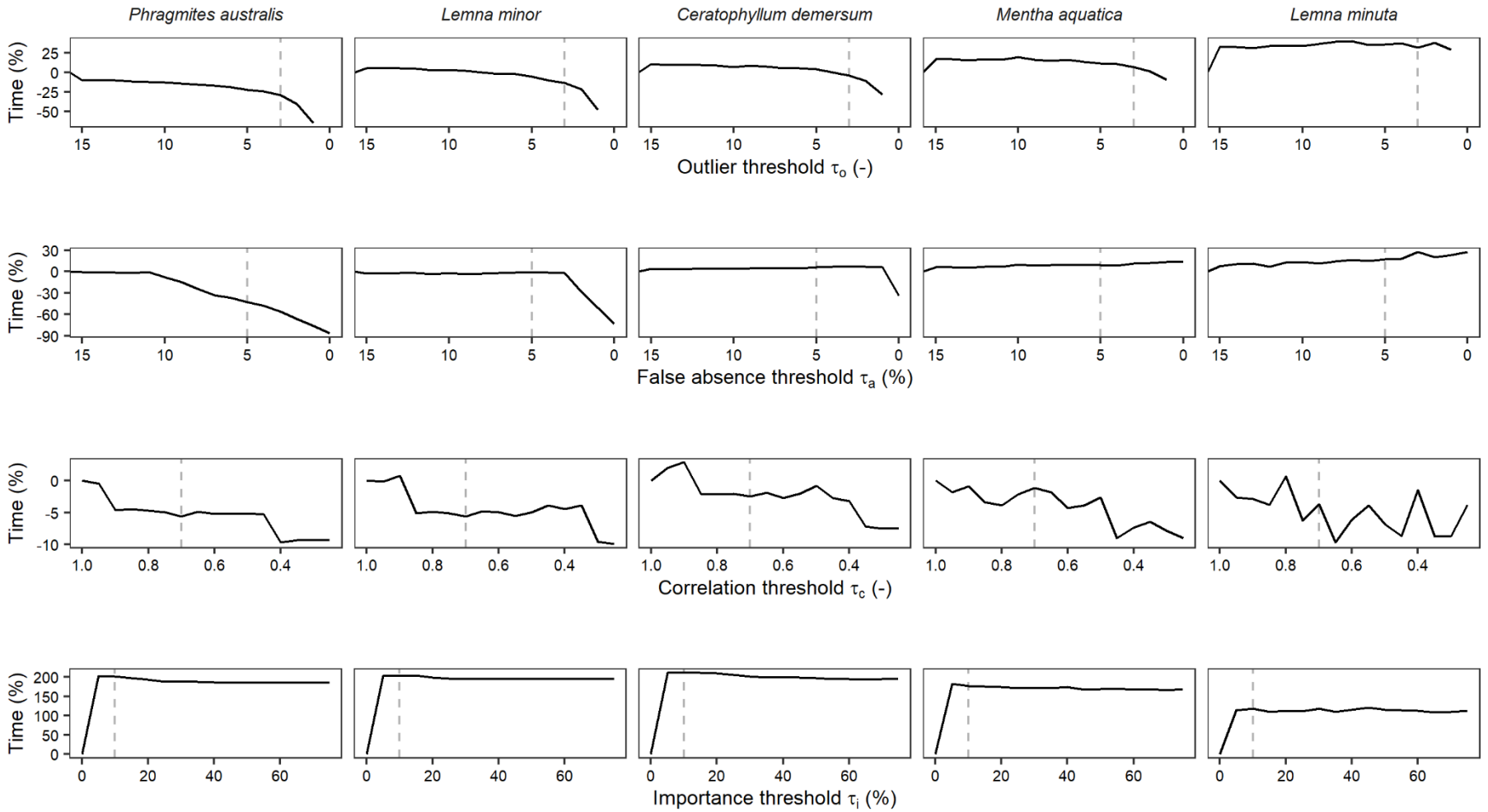

Figure C.6: Overview of four different pre-processing approaches and their effect on time required for sequential data pre-processing and model development. *In general, a positive effect of data pre-processing on overall computation time can be observed, though the effect depends on data availability. Only importance-based variable selection causes a clear increase in required computation time, mainly due to the fact of having to develop an additional model to derive variable importance scores.*

C.3 Threshold selection for all species

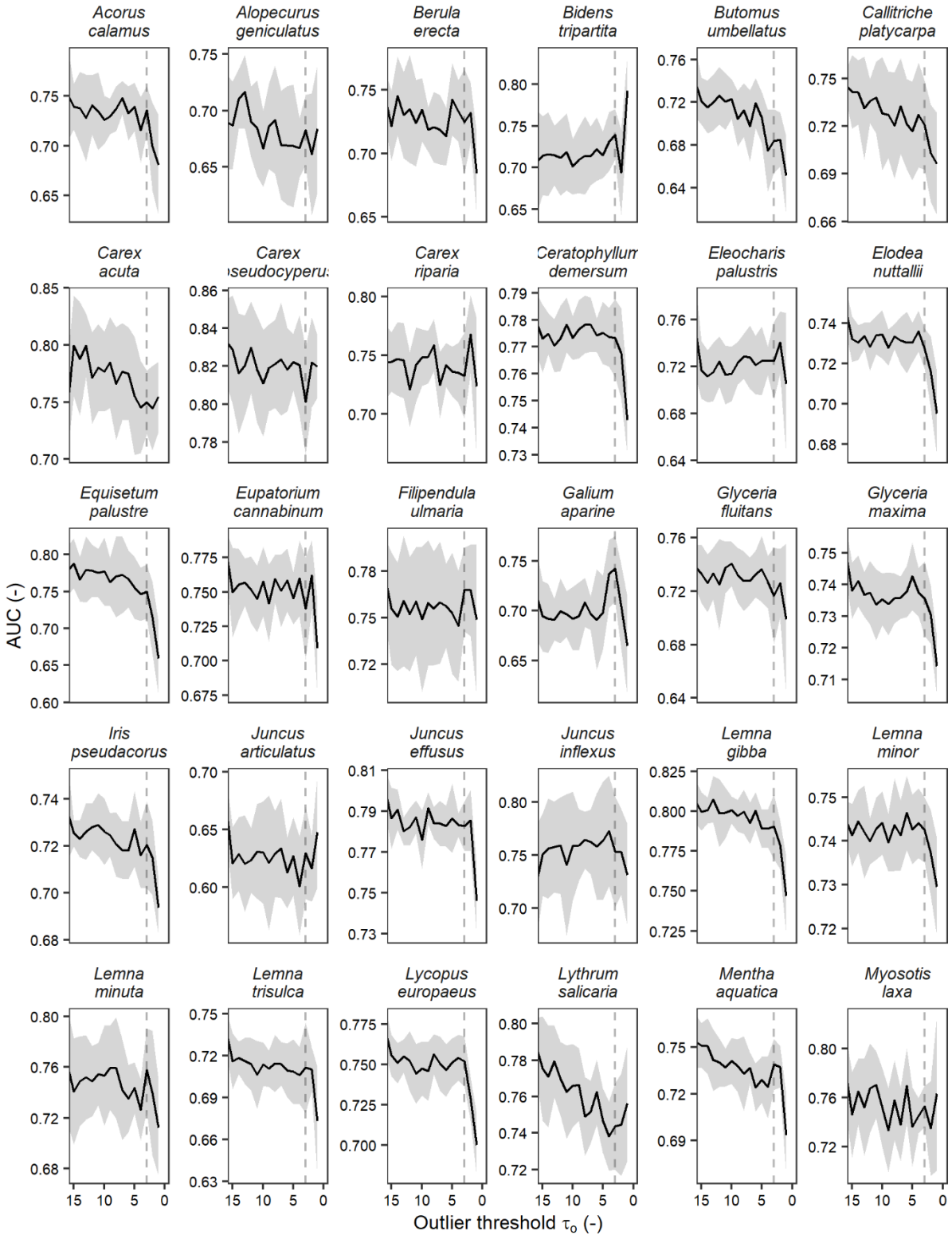

Figure C.7: Effect of outlier threshold selection on final model performance. *Analyses were performed for 58 different macrophyte species (see also Figure C.8) and illustrate the effect of outlier threshold selection (τ_o, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_o = 3$ (dashed grey line).*

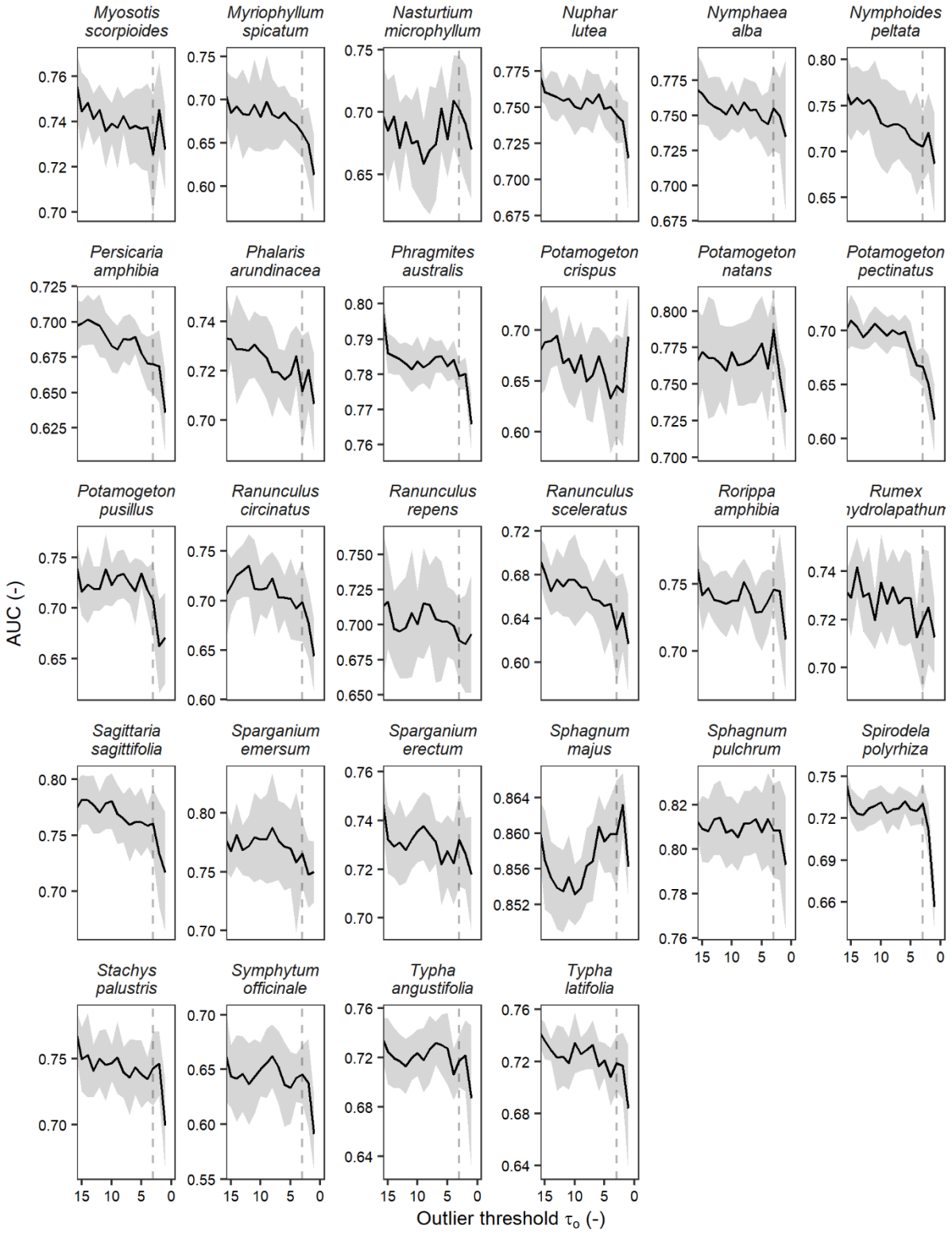

Figure C.8: Effect of outlier threshold selection on final model performance (continued). *Analyses were performed for 58 different macrophyte species (see also Figure C.7) and illustrate the effect of outlier threshold selection (τ_o, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_o = 3$ (dashed grey line).*

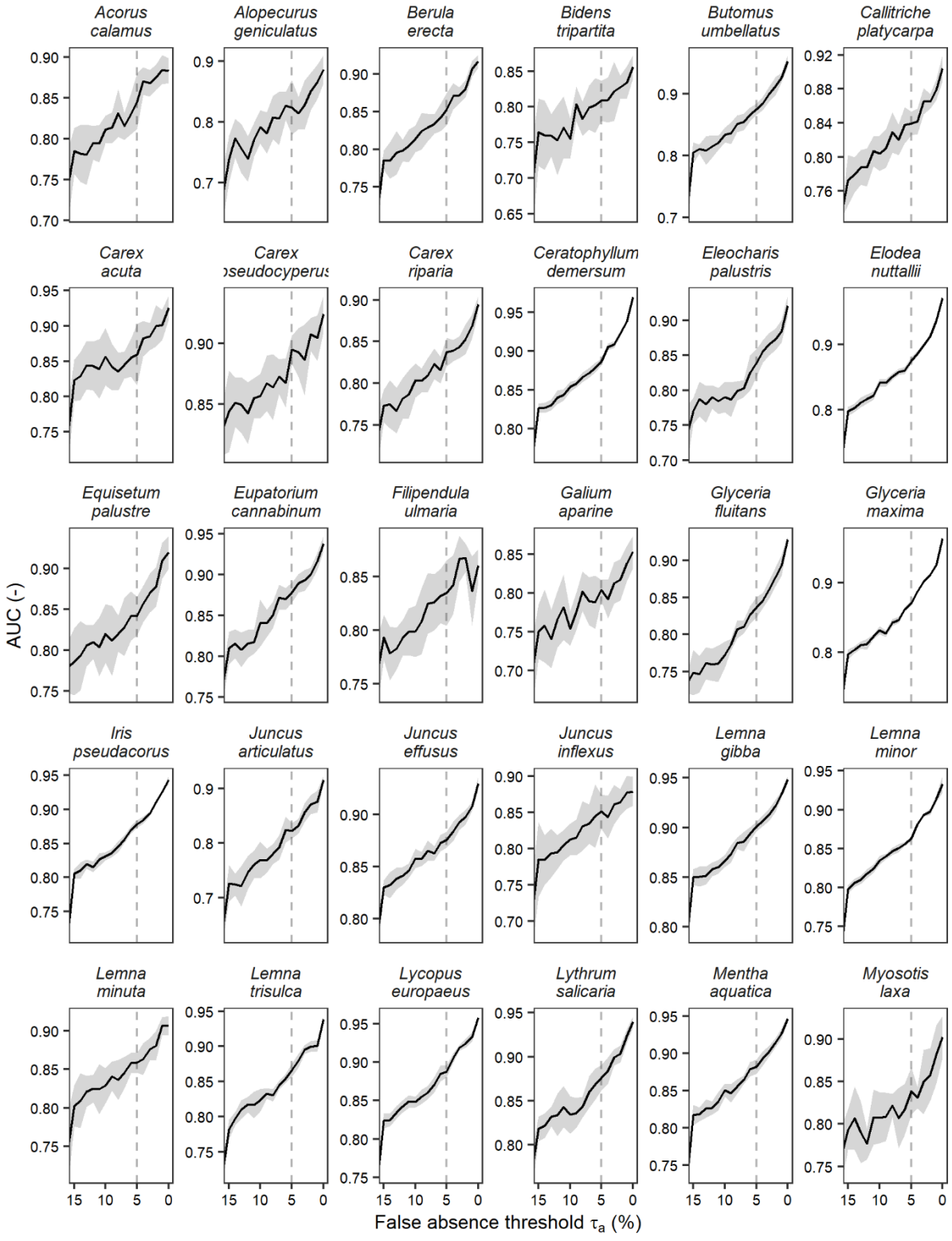

Figure C.9: Effect of false absences threshold selection on final model performance. *Analyses were performed for 58 different macrophyte species (see also Figure C.10) and illustrate the effect of outlier threshold selection (τ_a, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_a = 5$ % (dashed grey line).*

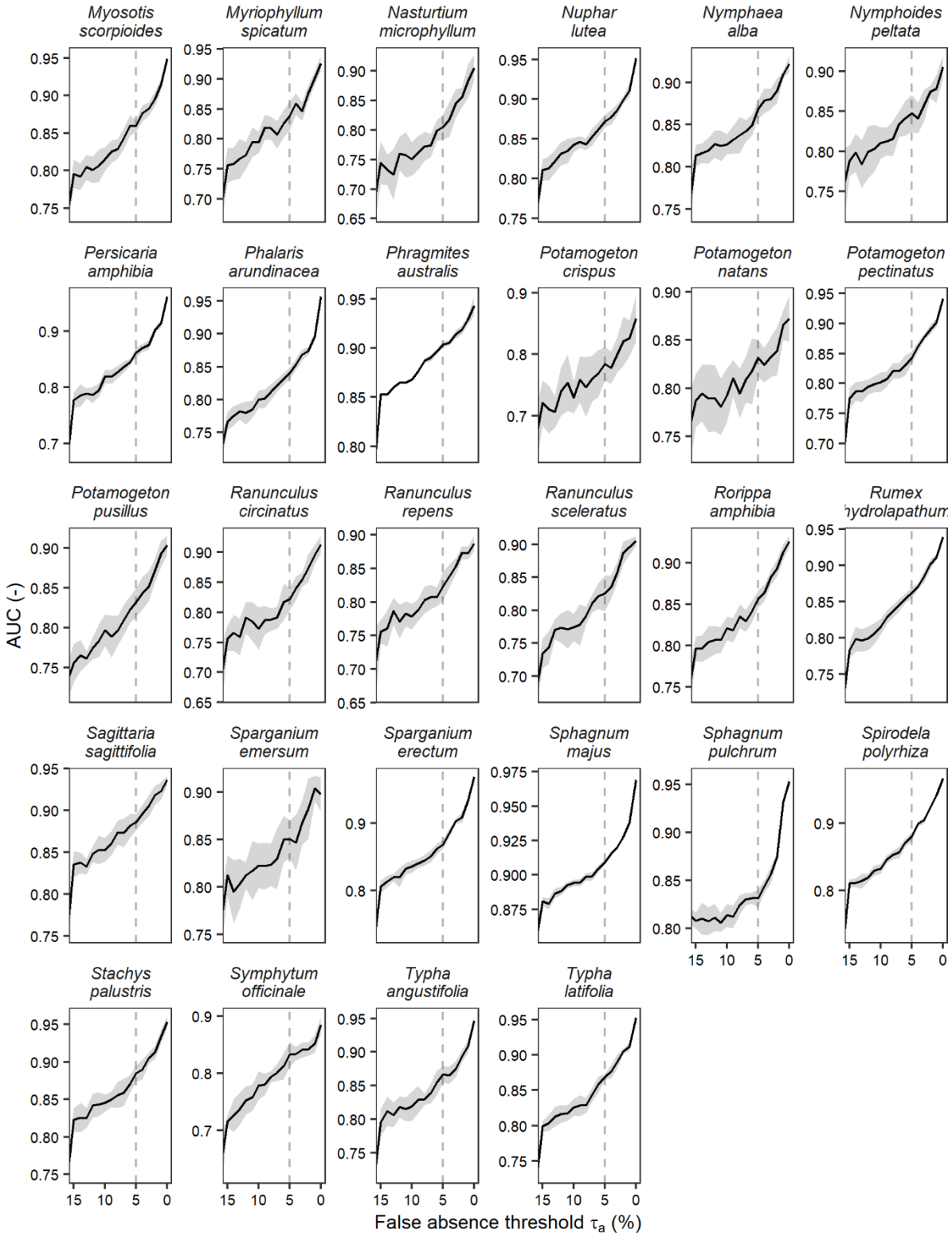

Figure C.10: Effect of false absences threshold selection on final model performance (continued). *Analyses were performed for 58 different macrophyte species (see also Figure C.9) and illustrate the effect of outlier threshold selection (τ_a, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_a = 5$ % (dashed grey line).*

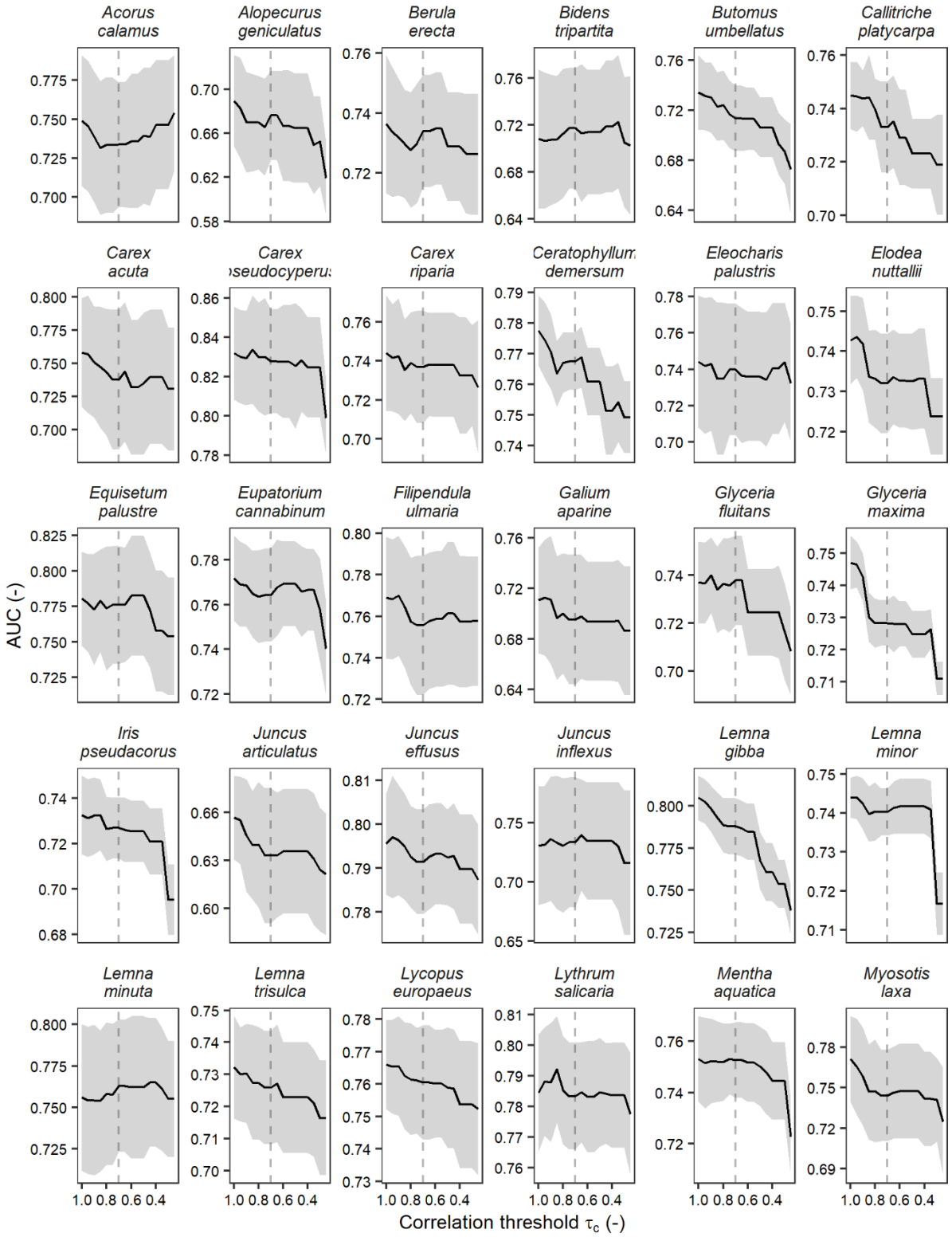

Figure C.11: Effect of correlation threshold selection on final model performance. *Analyses were performed for 58 different macrophyte species (see also Figure C.12) and illustrate the effect of outlier threshold selection (τ_c, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_c = 0.7$ (dashed grey line).*

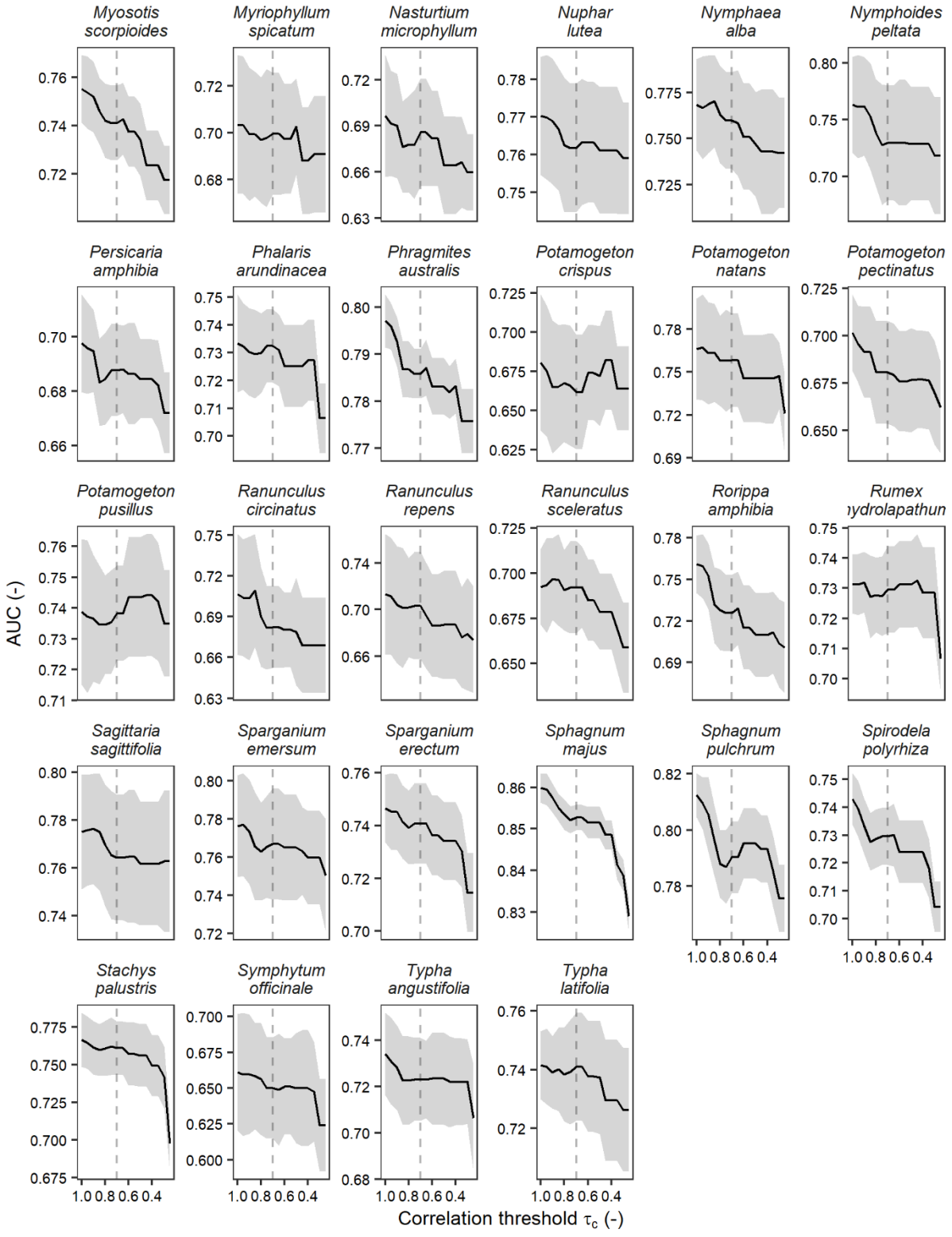

Figure C.12: Effect of correlation threshold selection on final model performance (continued). *Analyses were performed for 58 different macrophyte species (see also Figure C.11) and illustrate the effect of outlier threshold selection (τ_c, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_c = 0.7$ (dashed grey line).*

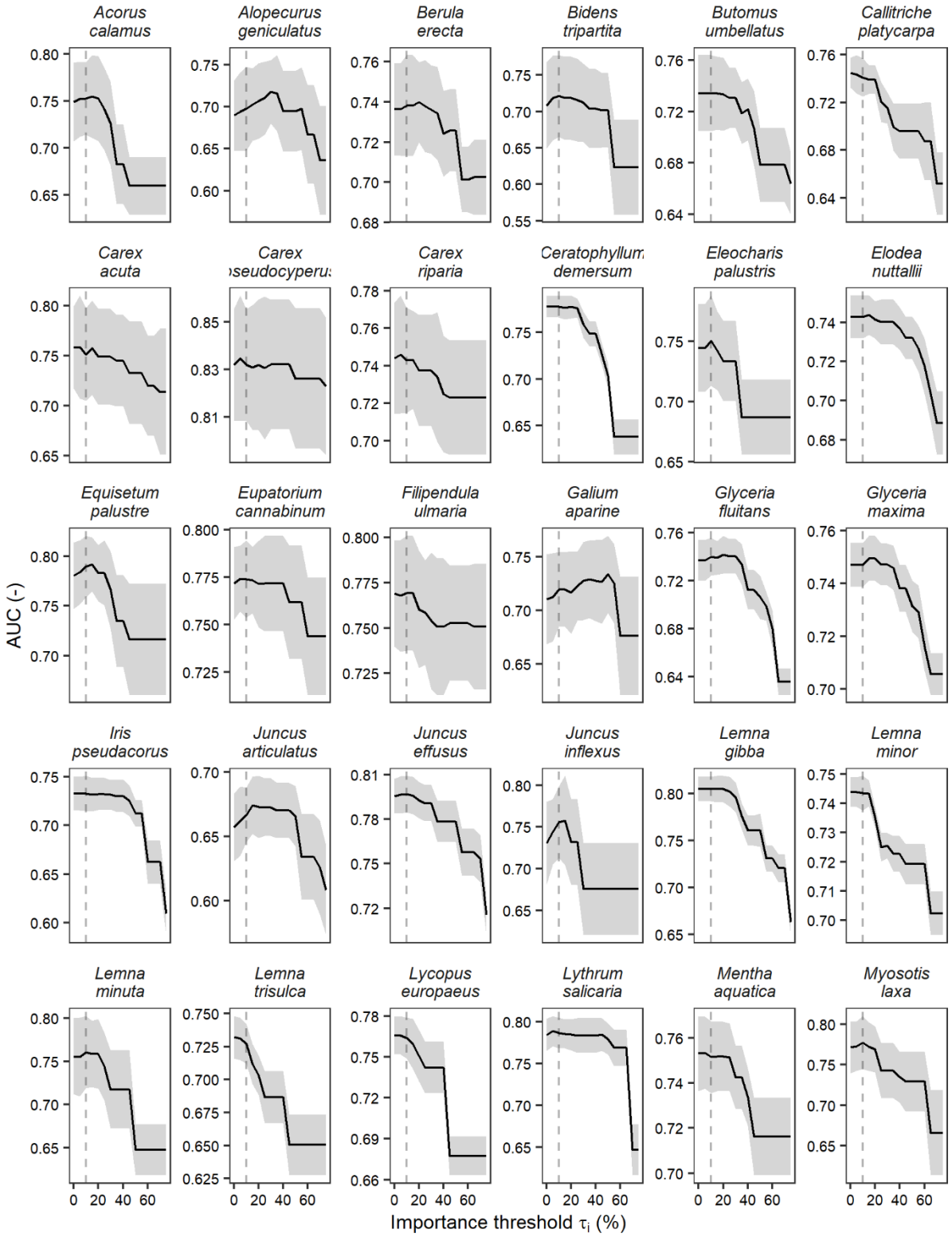

Figure C.13: Effect of importance threshold selection on final model performance. *Analyses were performed for 58 different macrophyte species (see also Figure C.14) and illustrate the effect of importance threshold selection (τ_i, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is $\tau_i = 10$ % (dashed grey line).*

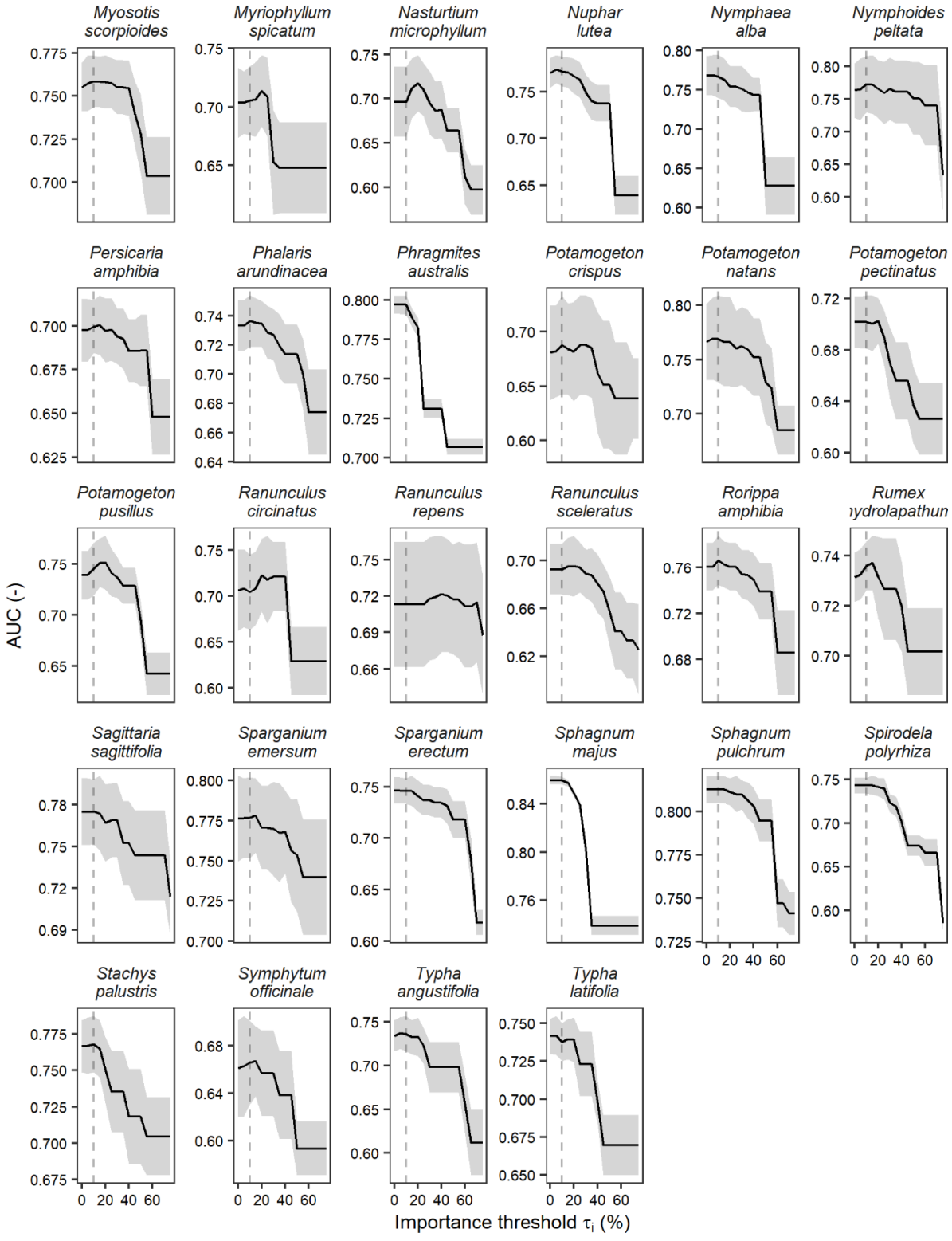

Figure C.14: Effect of importance threshold selection on final model performance (continued). *Analyses were performed for 58 different macrophyte species (see also Figure C.13) and illustrate the effect of importance threshold selection (τ_i, x-axis) on the discrimination performance of species-specific random forests (AUC, y-axis). Several patterns are obtained and indicate the potential of species-specific thresholds. The selected threshold in this work is τ_i = 10 % (dashed grey line).*

C.4 Environmental domains post-processing

Preferably, observational data that is used for the development of correlative habitat suitability models reflects the complete environmental domain, with presences occupying only a fraction of that domain. This allows for a distinction between suitable and unsuitable conditions within the final model, though is often challenged by data availability and sampling bias. The overlap between the occupied and observed environmental domain can be described at variable level or with a single metric, yet both techniques do not allow for a clear interpretation of actual domain overlap. On the one hand, variable-specific descriptions can find species presences at the lowest and highest observed variable values, which indicates that the considered variable does not cause a physiological limitation on the species' occurrence within its observed range. Absences observed at intermediate levels, however, can be caused by other variables exceeding the species' tolerance level, which indicates that the observed environmental domain exceeds the occupied environmental domain. On the other hand, distance metrics can help to summarise how far presences and absences are located from the centroid of the observed environmental domain. Presences can be expected to be located closer to the centroid and show less discrepancy or spread in the obtained distances, while absences extend the environmental domain defined by presences and are expected to show higher distance scores. However, it remains possible that an assumed absence is closely located to the centroid for all variables except one, with the exceptional variable causing the species to be absent. The resulting distance score can therefore be smaller than for a confirmed presence with overall deviating variable scores. Both analyses can help to create an impression of the domain overlap, though none provides a clear and unambiguous answer. This is illustrated with analyses performed for a selection of five macrophytes in Table C.3 and Figure C.15.

Table C.3: Overview of variable-specific ranges for a selection of macrophytes. *Ranges are reported as representing the observed environmental domain (environmental range; ER) and the occupied domain (species range; SR). The selected macrophytes align with the species reported within the main text of Chapter 6.*

Variable (unit)	Range	*P. australis*	*L. minor*	*C. demersum*	*M. aquatica*	*L. minuta*
Temperature (°C)	ER	3.5 – 29.6	3.5 – 29.6	3.5 – 29.6	3.5 – 29.6	3.5 – 29.6
	SR	5.5 – 28.6	4.5 – 29.0	7.1 – 29.0	8.6 – 27.4	13.0 – 27.6
pH (-)	ER	5.3 – 10.7	5.3 – 10.1	5.3 – 10.1	5.3 – 10.1	
	SR	5.3 – 10.0	5.3 – 10.0	6.3 – 9.9	6.0 – 9.6	
Chloride ($mg \cdot L^{-1}$)	ER	5.0 – 565			5.0 – 565	
	SR	7.0 – 560			10.0 – 510	
Oxygen ($mg \cdot L^{-1}$)	ER	0.0 – 21.2	0.0 – 21.2	0.0 – 21.2		0.0 – 21.2
	SR	0.0 – 20.8	0.0 – 21.2	0.0 – 20.6		1.1 – 17.0
Oxygen saturation (%)	ER				0.0 – 230	
	SR				0.0 – 200	
Transparency (m)	ER	0.0 – 1.7	0.0 – 1.7	0.0 – 1.7		
	SR	0.0 – 1.6	0.0 – 1.6	0.1 – 1.6		
Ammonium-N ($mg \cdot L^{-1}$)	ER	0.001 – 1.50	0.001 – 1.50	0.001 – 1.50	0.001 – 1.50	
	SR	0.01 – 1.50	0.01 – 1.50	0.01 – 1.50	0.01 – 1.42	
Nitrate-N ($mg \cdot L^{-1}$)	ER	0.01 – 7.30	0.01 – 7.30	0.01 – 7.30	0.01 – 7.30	0.01 – 7.30
	SR	0.01 – 7.00	0.01 – 7.20	0.01 – 7.15	0.01 – 7.00	0.04 – 3.30
Calcium ($mg \cdot L^{-1}$)	ER	0.04 – 200.0			0.04 – 200.0	
	SR	9.5 – 200.0			15.0 – 150.0	
Kjeldahl-N ($mg \cdot L^{-1}$)	ER		0.11 – 5.70	0.11 – 5.70		0.11 – 5.70
	SR		0.11 – 5.70	0.14 – 5.60		0.31 – 2.96

(Continues on next page)

(Continued)

Variable (unit)	Range	*P. australis*	*L. minor*	*C. demersum*	*M. aquatica*	*L. minuta*
Potassium ($mg \cdot L^{-1}$)	ER		0.11 – 45.0			
	SR		0.12 – 23.5			
Chlorophyll *a* ($\mu g \cdot L^{-1}$)	ER		0.0 – 158.3	0.0 – 158.3		0.0 – 158.3
	SR		1.0 – 158.3	1.0 – 145.3		5.0 – 72.9
Total Phosphorus ($mg \cdot L^{-1}$)	ER			0.01 – 1.6	0.01 – 1.6	
	SR			0.01 – 1.3	0.01 – 1.1	
Sulphate ($mg \cdot L^{-1}$)	ER			1.0 – 310		1.0 – 310
	SR			6.0 – 310		10.6 – 138.0
BOD5 ($mg \cdot L^{-1}$)	ER				0.0 – 13.0	
	SR				1.0 – 13.0	

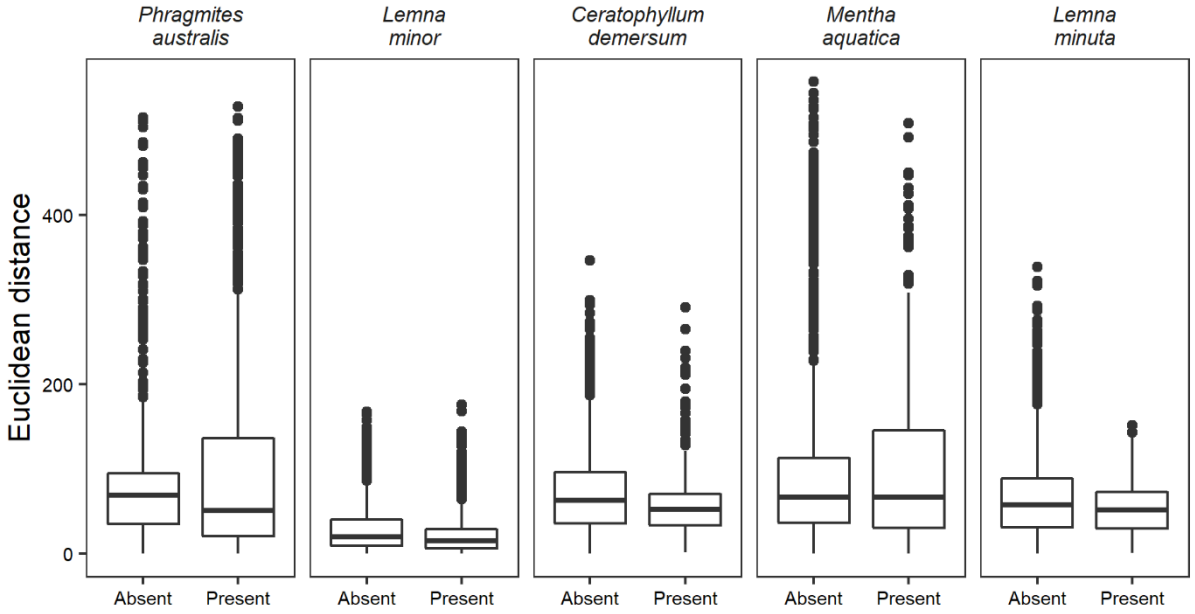

Figure C.15: Distribution of distance metrics for locations with and without species occurrence for a selection of five macrophytes. *Distances are calculated as the Euclidean distance between the environmental conditions at a specific location and the centroid of the observed environmental domain. The selected macrophytes align with the species reported within the main text of Chapter 6. Boxes represent the 50 % central values around the median, while whiskers represent the first and third quartile extended to the last case within 1.5 times the interquartile range. Dots represent the values outside the range of the whiskers.*

D

Supportive Information for Chapter 7 – Developing abiotic habitat suitability models

D.1 Data characteristics

Table D.1: Characteristics of the species-specific data sets after data pre-processing. *For each species, the original data (see Figure 4.5) was subjected to outlier, false absence, correlated and irrelevant variable removal. All methods, except outlier removal, are species-specific and result in different data set characteristics. An overview of the specific variables being included for each species can be found in Figure D.1.*

Macrophyte	Instances	Variables	Prevalence (%)
Acorus calamus	1958	5	7
Alopecurus geniculatus	1882	6	4.14
Berula erecta	2016	11	9.03
Bidens tripartita	2055	5	4.38
Butomus umbellatus	2101	10	10.47
Callitriche platycarpa	1804	9	10.25
Carex acuta	2368	3	3.63
Carex pseudocyperus	2411	6	4.19
Carex riparia	1998	9	7.61
Ceratophyllum demersum	2314	10	28.95
Eleocharis palustris	1725	7	9.1
Elodea nuttallii	2046	10	37.29
Equisetum palustre	2136	3	4.12
Eupatorium cannabinum	2142	10	8.4
Filipendula ulmaria	2141	5	5.04
Galium aparine	1991	6	4.17
Glyceria fluitans	1446	11	18.19
Glyceria maxima	2147	10	46.86
Iris pseudacorus	2075	9	31.86
Juncus articulatus	1846	9	7.91
Juncus effusus	1815	8	24.02
Juncus inflexus	2637	7	3.49
Lemna gibba	2083	9	17.52
Lemna minor	2193	9	43.64
Lemna minuta	2398	6	5.46
Lemna trisulca	2032	7	17.57
Lycopus europaeus	2064	7	23.21
Lythrum salicaria	1914	6	12.33
Mentha aquatica	2083	9	18.05
Myosotis laxa	2078	7	4.86
Myosotis scorpioides	1898	10	16.23

(Continues on next page)

(Continued)

Macrophyte	Instances	Variables	Prevalence (%)
Myriophyllum spicatum	2007	8	7.67
Nasturtium microphyllum	1879	10	5.32
Nuphar lutea	2089	9	18.81
Nymphaea alba	2193	7	10.9
Nymphoides peltata	2089	8	5.46
Persicaria amphibia	1897	10	22.14
Phalaris arundinacea	1593	8	25.3
Phragmites australis	2407	8	55.38
Potamogeton crispus	1559	10	5.2
Potamogeton natans	2079	11	6.06
Potamogeton pectinatus	1929	11	14.15
Potamogeton pusillus	1798	8	9.84
Ranunculus circinatus	2300	7	3.78
Ranunculus repens	1570	10	10.76
Ranunculus sceleratus	1716	10	10.9
Rorippa amphibia	1979	8	11.27
Rumex hydrolapathum	1904	10	21.27
Sagittaria sagittifolia	2122	5	10.98
Sparganium emersum	2075	8	5.98
Sparganium erectum	2093	9	23.94
Sphagnum majus	2399	9	42.1
Sphagnum pulchrum	1688	8	22.81
Spirodela polyrhiza	2112	10	30.87
Stachys palustris	2150	9	11.86
Symphytum officinale	1941	8	8.5
Typha angustifolia	2062	7	11.3
Typha latifolia	2122	7	19.79

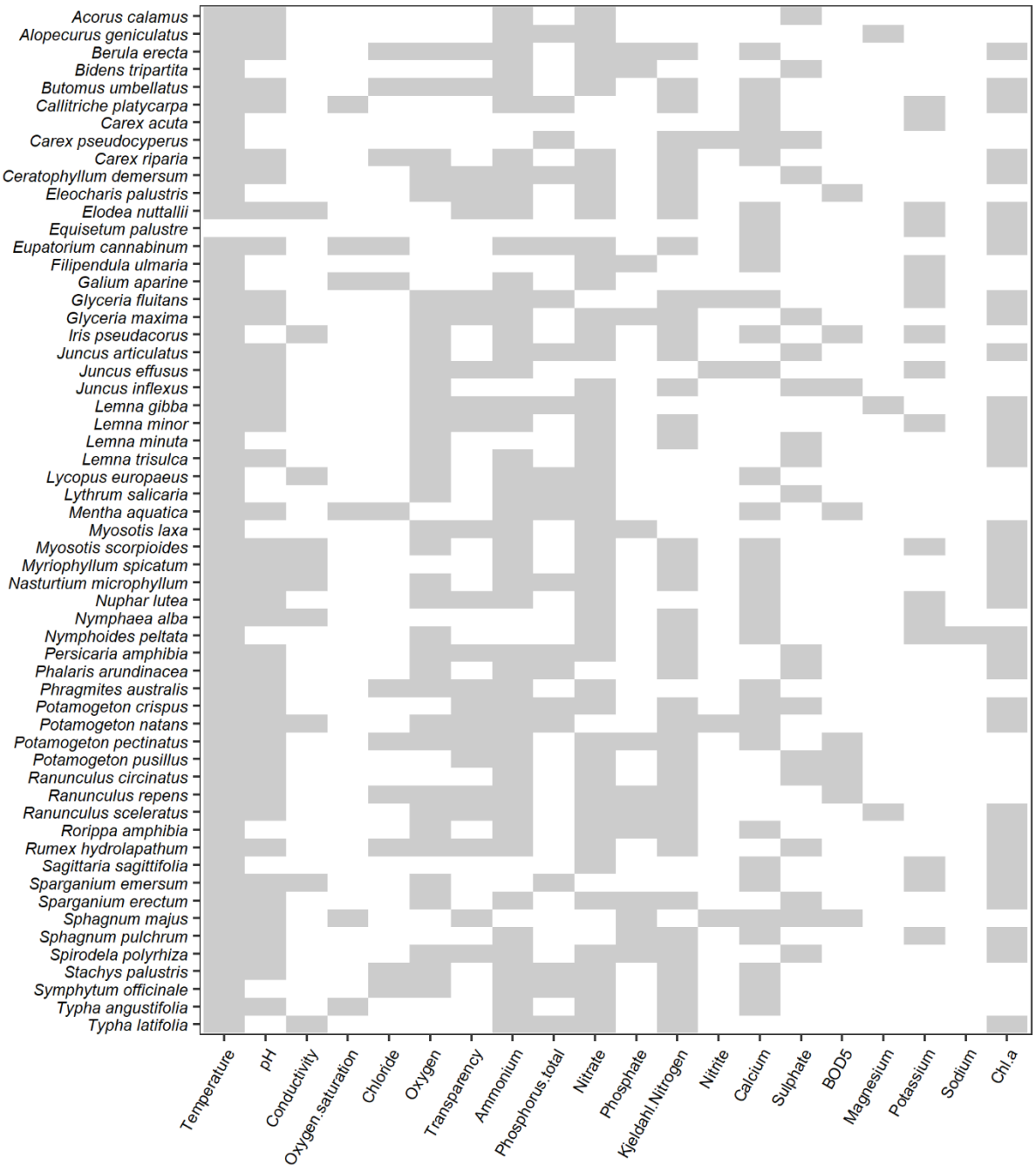

Figure D.1: Variable inclusion in species-specific training data. *For each of the selected 58 macrophyte species, individual data pre-processing was implemented, leading to different variables being included in the final training data. The number of included variables ranged from 3 up to 11 (see also Table D.1), as indicated by the grey cells. White cells depict variables that were not included in the species-specific training data.*

D.2 Variable importance

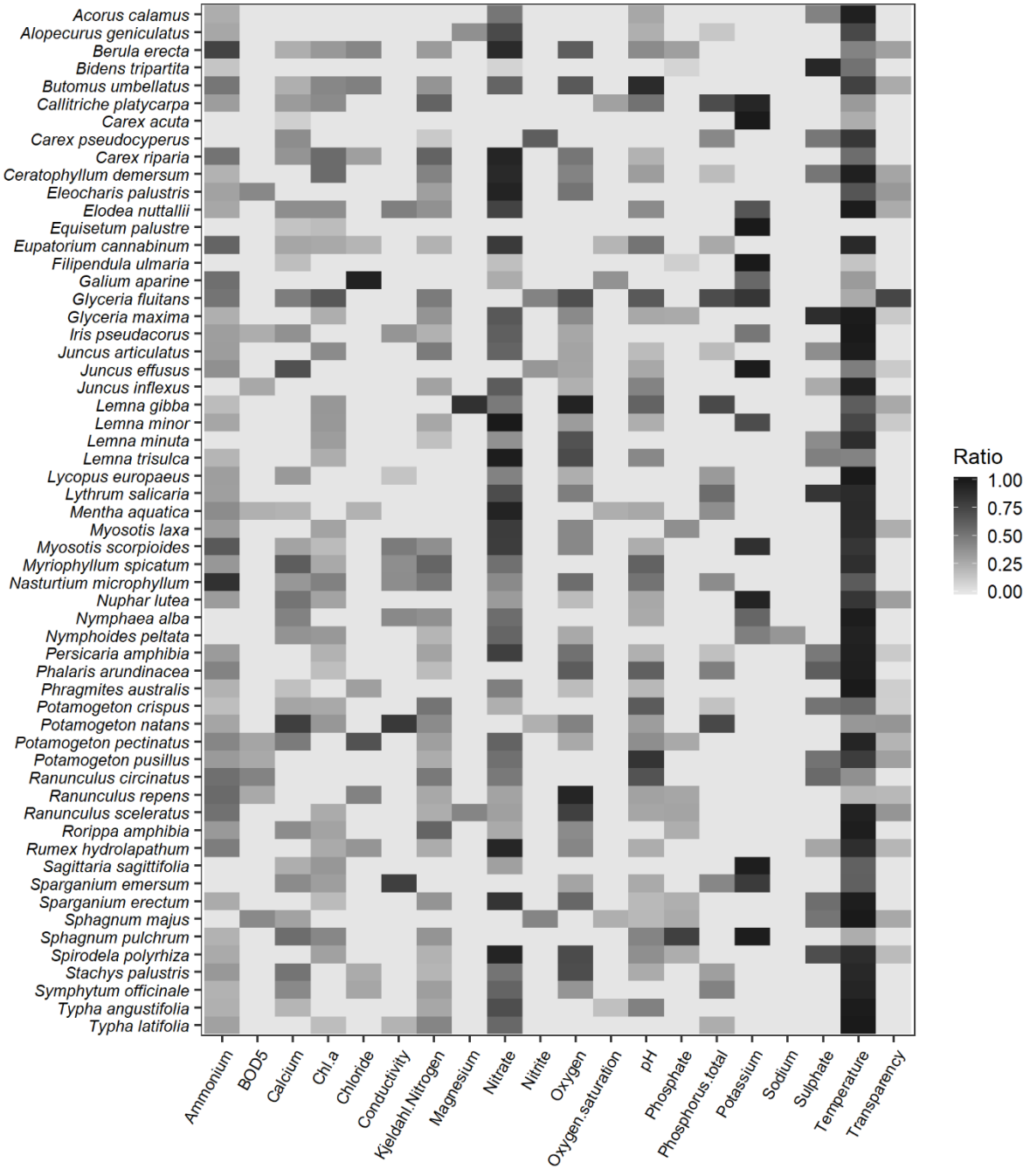

Figure D.2: Heatmap of considered and important variable for each macrophyte. *Scores range between 0 (light grey) and 1 (black) and reflect the model improvement ratio (MIR) over 10 repetitions of 5-fold cross-validation, with higher scores representing a higher relative importance of the variable. Temperature is considered an important variable for most macrophytes as is nitrate. Ammonium, oxygen and pH are present in most models with intermediate MIR scores.*

D.3 Scenario analysis

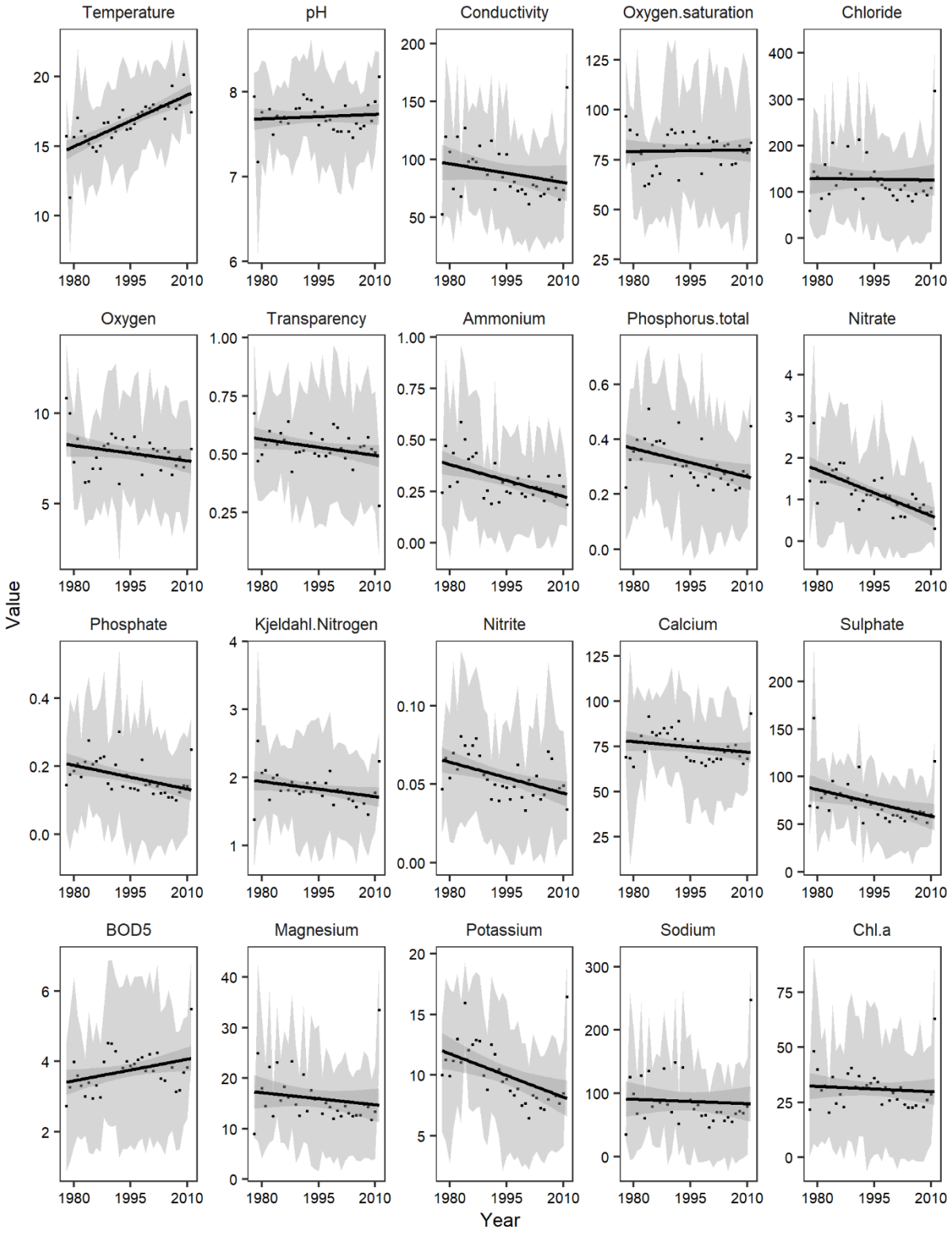

Figure D.3: Temporal patterns in abiotic data used for model development. *Dots indicate the annual averages (April to September) with light grey ribbons covering the standard deviation. Black solid lines represent the temporal trends, complemented with a dark grey confidence interval. The latter is relatively small compared to the uncertainty on the annual averages. Quantitative expression of variable-specific intercepts and slopes can be found in Table D.2.*

Table D.2: Variable-specific summary of average conditions in 2010 and linear models fitted to the temporal data. *For each variable, the mean and standard deviation (sd) are calculated for the months April until September and rounded to two digits, along with an intercept and coefficient (for time) and supplemented with their 95 % confidence intervals (CI_{95}). Model fitting was based on training data from 58 macrophyte species. Graphical representation of linear models is shown in Figure D.3. *: reported value in the range [-0.001; 0.001].*

Variable	Unit	Mean	Sd	Intercept		Coefficient	
				Value	CI_{95}	Value	CI_{95}
Temperature	°C	18.7	2.8	-260.2	[-283.1; -237.3]	0.139	[0.127; 0.150]
pH	-	7.9	0.6	-0.91	[-4.93; 3.11]	0.004	[0.002; 0.006]
Conductivity	$mS \cdot m^{-1}$	75.25	43.41	1150.6	[754.8; 1546.5]	-0.533	[-0.731; -0.335]
Oxygen saturation	%	78.75	42.21	-322.8	[-592.3; -53.2]	0.201	[0.066; 0.336]
Chloride	$mg \cdot L^{-1}$	112.76	99.05	108.2	[-780.2; 997.3]	0.008	[-0.437; 0.452]
Oxygen	$mg \cdot L^{-1}$	7.41	3.34	31.34	[7.60; 55.07]	-0.012	[-0.024; 0.000*]
Transparency	m	0.50	0.26	6.55	[4.65; 8.44]	-0.003	[-0.004; -0.002]
Ammonium-N	$mg \cdot L^{-1}$	0.27	0.19	9.68	[7.71; 11.64]	-0.005	[-0.006; -0.004]
Phosphorus total	$mg \cdot L^{-1}$	0.28	0.19	7.06	[5.32; 8.79]	-0.003	[-0.004; -0.003]
Nitrate-N	$mg \cdot L^{-1}$	0.67	0.72	77.54	[68.77; 86.31]	-0.038	[-0.043; -0.034]
Phosphate-P	$mg \cdot L^{-1}$	0.14	0.14	4.50	[3.28; 5.73]	-0.002	[-0.003; -0.002]
Kjeldahl nitrogen-N	$mg \cdot L^{-1}$	1.79	0.57	13.27	[7.97; 18.57]	-0.006	[-0.008; -0.003]
Nitrite-N	$mg \cdot L^{-1}$	0.05	0.03	1.77	[1.45; 2.09]	-0.001	[-0.001; -0.001]
Calcium	$mg \cdot L^{-1}$	68.78	17.95	748.7	[563.1; 934.4]	-0.338	[-0.431; -0.245]
Sulphate	$mg \cdot L^{-1}$	61.37	30.92	1322.0	[1044.4; 1600.0]	-0.627	[-0.766; -0.488]
BOD5	$mg \cdot L^{-1}$	3.89	1.50	-37.71	[-51.69; -23.73]	0.021	[0.014; 0.028]
Magnesium	$mg \cdot L^{-1}$	14.11	10.09	91.84	[-6.67; 190.34]	-0.038	[-0.088; 0.011]
Potassium	$mg \cdot L^{-1}$	8.68	4.51	215.2	[175.3; 255.0]	-0.103	[-0.123; -0.083]
Sodium	$mg \cdot L^{-1}$	85.37	82.30	-64.09	[-916.0; 787.8]	0.074	[-0.352; 0.501]
Chlorophyll *a*	$\mu g \cdot L^{-1}$	29.49	23.57	-42.66	[-246.44; 161.13]	0.036	[-0.065; 0.138]

Table D.3: Start and end points for the developed scenarios. *Scenarios are defined in Table 7.2, with end points defined via linear regression (see Table D.2). A selection of endpoints in the KEY scenarios were reached via exponential patterns instead of linear patterns (indicated with #).*

Variable	Unit	AVG			EXT			NUT		
		Start	BAU	KEY	Start	BAU	KEY	Start	BAU	KEY
Temperature	°C	18.66	22.02	22.02	18.66	22.02	22.02	18.66	22.02	22.02
pH	-	7.892	8.027	8.027	7.892	8.027	8.027	7.892	8.027	8.027
Conductivity	$mS \cdot m^{-1}$	75.247	66.993	66.993	162.076	153.822	153.822	75.247	66.993	66.993
Oxygen saturation	%	78.747	85.465	85.465	36.534	43.251	43.251	78.747	85.465	85.465
Chloride	$mg \cdot L^{-1}$	112.759	117.958	117.958	310.851	316.050	316.050	112.759	117.958	117.958
Oxygen	$mg \cdot L^{-1}$	7.405	7.346	5.000	4.065	4.007	5.000	7.405	7.346	5.000
Transparency	m	0.502	0.436	0.436	0.502	0.436	0.436	0.502	0.436	0.436
Ammonium-N	$mg \cdot L^{-1}$	0.271	0.112	0.200#	0.641	0.483	0.200#	0.641	0.483	0.200#
Phosphorus total	$mg \cdot L^{-1}$	0.275	0.186	0.186	0.657	0.568	0.568	0.657	0.568	0.568
Nitrate-N	$mg \cdot L^{-1}$	0.695	0.001	0.500#	2.126	1.165	0.500#	2.126	1.165	0.500#
Phosphate-P	$mg \cdot L^{-1}$	0.143	0.084	0.084	0.423	0.367	0.367	0.426	0.367	0.367
Kjeldahl nitrogen-N	$mg \cdot L^{-1}$	1.788	1.652	1.652	2.920	2.784	2.784	2.920	2.784	2.784
Nitrite-N	$mg \cdot L^{-1}$	0.049	0.024	0.024	0.117	0.092	0.092	0.117	0.092	0.092
Calcium	$mg \cdot L^{-1}$	68.780	61.050	61.050	104.684	96.953	96.953	68.780	61.050	61.050
Sulphate	$mg \cdot L^{-1}$	61.373	47.668	47.668	123.218	109.513	109.513	61.373	47.668	47.668
BOD5	$mg \cdot L^{-1}$	3.885	4.451	4.450	6.877	7.442	7.442	3.885	4.450	4.450
Magnesium	$mg \cdot L^{-1}$	14.106	13.752	13.752	34.278	33.924	33.924	14.106	13.752	13.752
Potassium	$mg \cdot L^{-1}$	8.684	6.192	6.192	17.703	15.210	15.210	8.684	6.192	6.192
Sodium	$mg \cdot L^{-1}$	85.371	90.123	90.123	249.971	254.724	254.724	85.371	90.123	90.123
Chlorophyll *a*	$\mu g \cdot L^{-1}$	29.485	32.315	32.315	76.621	79.451	79.451	29.485	32.315	32.315

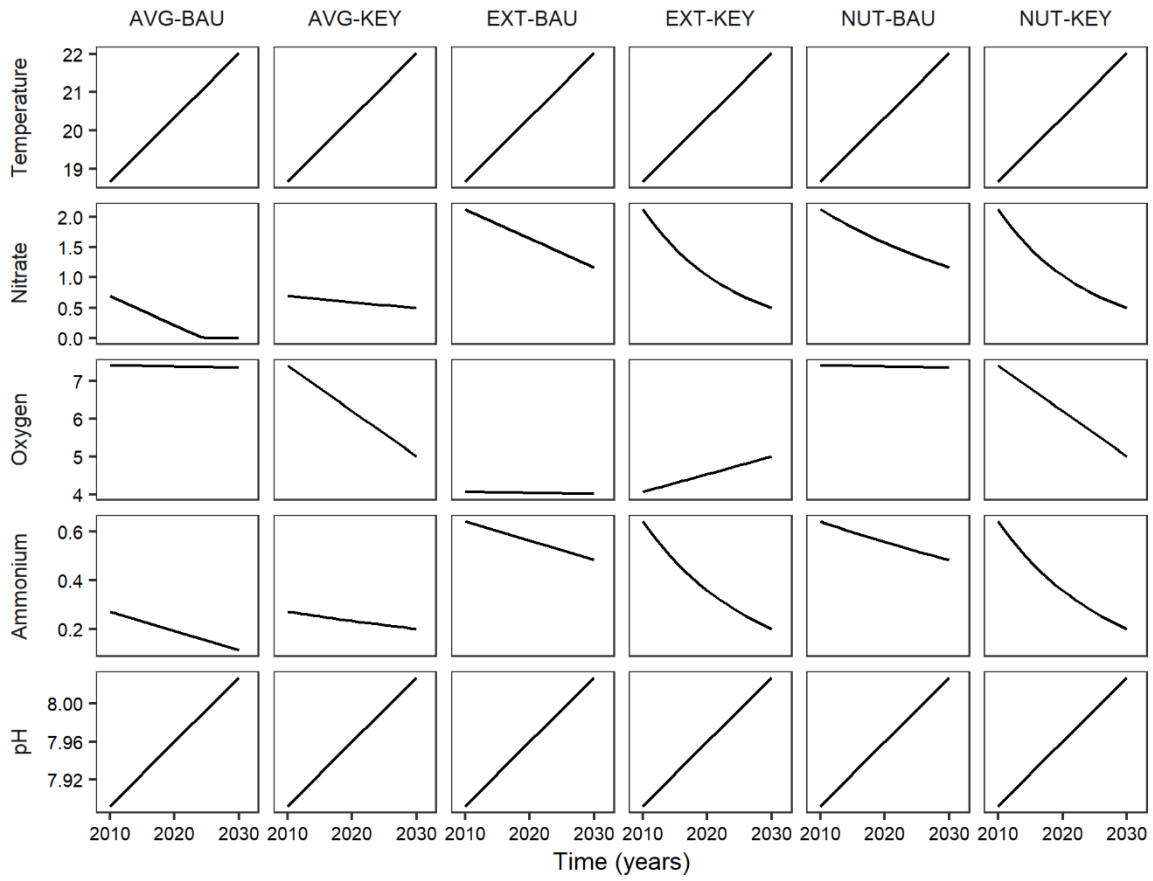

Figure D.4: Depiction of the different scenarios for the five most steering variables for a period of 20 years. *Starting points (AVG, EXT and NUT) were based on the average conditions in 2010 (see Table D.2) with specific differences among AVG (general mean), EXT ($\bar{\boldsymbol{x}} + \mathbf{2} \cdot \boldsymbol{s}$) and NUT ($\bar{\boldsymbol{x}} + \mathbf{2} \cdot \boldsymbol{s}$ for nutrients). Management consisted of business-as-usual (BAU) and relied on the inferred temporal linear models (see Table D.2 and Figure D.3), while separate focus on key variables (KEY) was based on reaching the optimal conditions inferred from the partial dependence plots (see Figure 7.2).*

D.4 Species-specific temporal trends

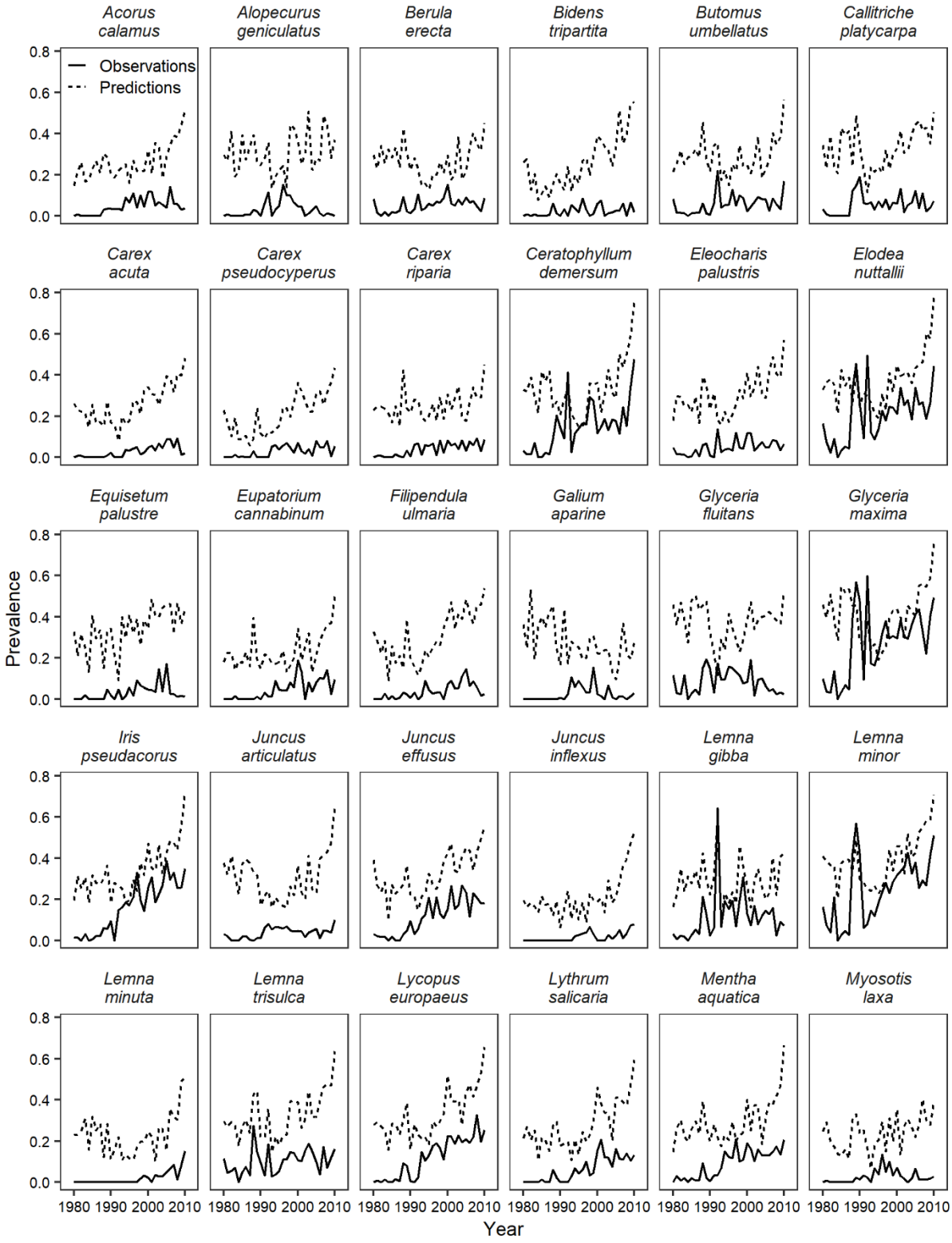

Figure D.5: Temporal trend of observed and predicted prevalence of all 58 macrophytes. *Prevalence is determined by the fraction of sites where macrophyte presence is observed (solid line) or where conditions are suitable to support macrophyte presence (dashed line). The fraction of both suitable and occupied sites increases in time and indicates a suboptimal use of the available suitable habitats.*

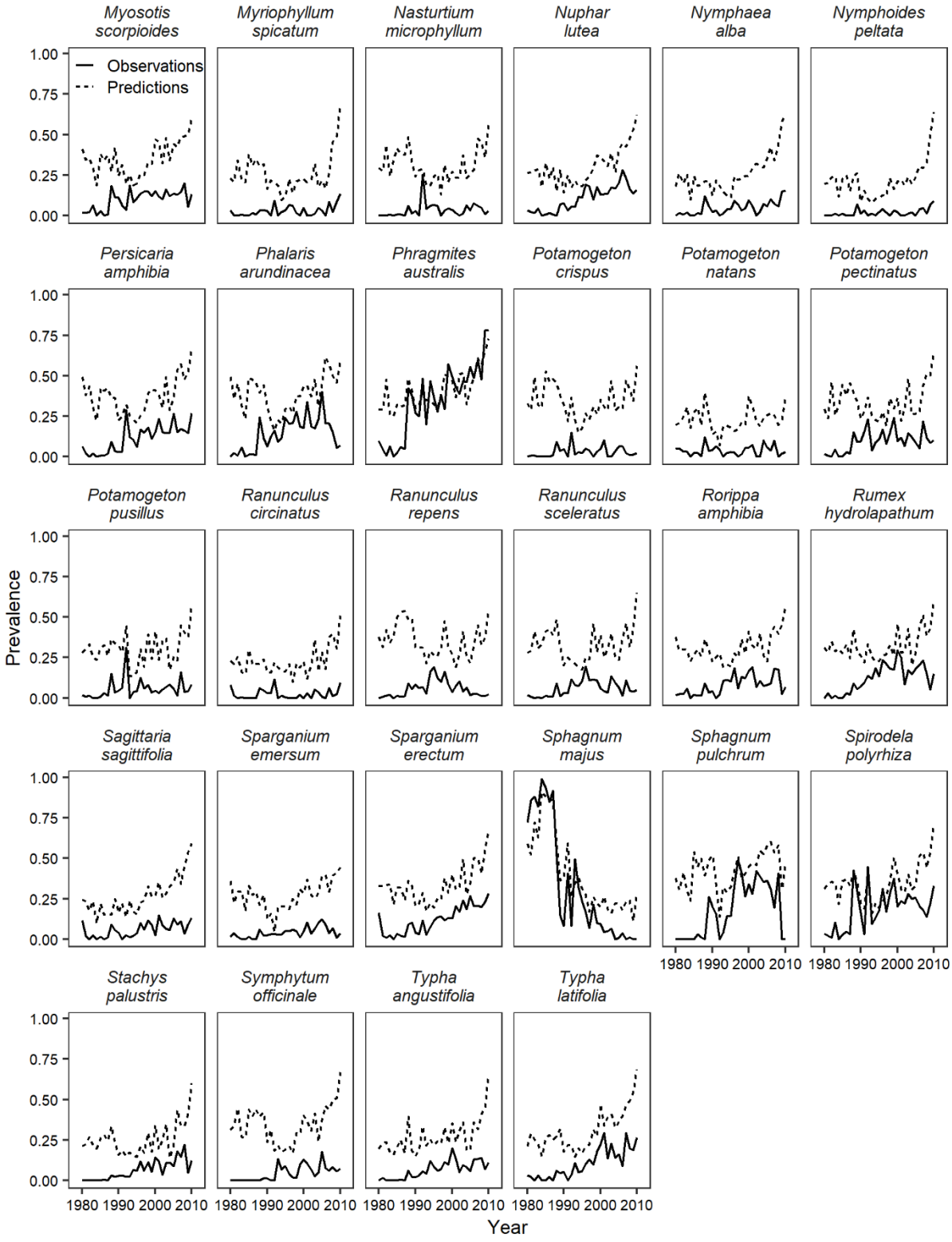

Figure D.6: Temporal trend of observed and predicted prevalence of all 58 macrophytes (continued). *Prevalence is determined by the fraction of sites where macrophyte presence is observed (solid line) or where conditions are suitable to support macrophyte presence (dashed line). The fraction of both suitable and occupied sites increases in time and indicates a suboptimal use of the available suitable habitats.*

E

Supportive Information for Chapter 8 – Functional traits for assessing invasive potential

E.1 Tables supporting results

Table E.1: Average total nitrogen (tN) concentration at day 0 and day 4 in mg·L^{-1}. *The average and standard deviation for each concentration is based on six separate samples.*

	Day 0	Day 4		
		Reference	*L. minor*	*L. minuta*
C1	70 *(± 2)*	70 *(± 3)*	62 *(± 2)*	62 *(± 2)*
C2	33 *(± 2)*	33 *(± 1)*	24 *(± 7)*	28 *(± 2)*
C3	16 *(± 2)*	15 *(± 1)*	14 *(± 6)*	12.7 *(± 0.7)*
C4	8.8 *(± 0.5)*	9.1 *(± 0.4)*	2.6 *(± 0.6)*	3.6 *(± 0.8)*
C5	4.2 *(± 0.1)*	4.6 *(± 0.4)*	0.4[a] *(± 0.6)*	1.0[a] *(± 0.6)*

[a] Contains samples with nitrogen concentration below detection limit.

Table E.2: Average total phosphorus (tP) concentration at day 0 and day 4 in mg·L^{-1}. *The average and standard deviation for each concentration is based on six separate samples.*

	Day 0	Day 4		
		Reference	*L. minor*	*L. minuta*
C1	20.99 *(± 0.09)*	20.2 *(± 0.5)*	18 *(± 1)*	17 *(± 1)*
C2	10.7 *(± 0.1)*	10.2 *(± 0.3)*	8 *(± 1)*	9 *(± 1)*
C3	5.43 *(± 0.07)*	5.3 *(± 0.6)*	3.4 *(± 0.3)*	4.0 *(± 0.5)*
C4	2.58 *(± 0.03)*	2.5 *(± 0.1)*	1.4 *(± 0.6)*	1.5 *(± 0.2)*
C5	1.33 *(± 0.01)*	1.23 *(± 0.08)*	0.4 *(± 0.1)*	0.5 *(± 0.1)*

***Table E.3: Evolution of the dry weight (in mg) during the first two days of the experiment for* L. minor *and* L. minuta.** *The average and standard deviation for each concentration is based on six separate samples.*

	L. minor		*L. minuta*	
	Day 0	Day 2	Day 0	Day 2
C1	23 *(± 2)*	18[a] *(± 6)*	20.7 *(± 0.8)*	27[a] *(± 5)*
C2	23 *(± 2)*	23[a] *(± 7)*	22 *(± 3)*	27 *(± 8)*
C3	23 *(± 1)*	25 *(± 7)*	20 *(± 1)*	33[a] *(± 3)*
C4	23 *(± 1)*	21 *(± 10)*	20.5 *(± 0.9)*	31[a] *(± 11)*
C5	22 *(± 1)*	20[a] *(± 8)*	22 *(± 1)*	22 *(± 11)*

[a] Dry weight content of one sample cannot be determined and is removed.

***Table E.4: Evolution of the dry weight (in mg) during the last two days of the experiment for* L. minor *and* L. minuta.** *The average and standard deviation for each concentration is based on six separate samples.*

	L. minor		*L. minuta*	
	Day 2	Day 4	Day 2	Day 4
C1	15[a] *(± 5)*	48 *(± 5)*	20[a] *(± 4)*	48 *(± 5)*
C2	18[a] *(± 4)*	49 *(± 4)*	19 *(± 5)*	52 *(± 5)*
C3	19 *(± 4)*	51 *(± 5)*	18[a] *(± 1)*	41 *(± 7)*
C4	16 *(± 5)*	46 *(± 6)*	17[a] *(± 2)*	36 *(± 6)*
C5	14[a] *(± 6)*	45 *(± 6)*	17 *(± 6)*	43 *(± 6)*

[a] Dry weight content of one sample cannot be determined and is removed.

***Table E.5: Evolution of the fresh weight (in mg) during the first two days of the experiment for* L. minor *and* L. minuta.** *The average and standard deviation for each concentration is based on six separate samples.*

	L. minor		*L. minuta*	
	Day 0	Day 2	Day 0	Day 2
C1	500 *(± 2)*	580 *(± 70)*	500 *(± 3)*	600 *(± 100)*
C2	499 *(± 2)*	640 *(± 70)*	499 *(± 2)*	710 *(± 60)*
C3	501 *(± 3)*	670 *(± 40)*	499 *(± 3)*	800 *(± 200)*
C4	500 *(± 2)*	700 *(± 100)*	500 *(± 2)*	800 *(± 200)*
C5	500 *(± 3)*	700 *(± 100)*	500 *(± 3)*	800 *(± 100)*

***Table E.6: Evolution of the fresh weight (in mg) during the last two days of the experiment for* L. minor *and* L. minuta.** *The average and standard deviation for each concentration is based on six separate samples.*

	L. minor		*L. minuta*	
	Day 2	Day 4	Day 2	Day 4
C1	490 *(± 20)*	710 *(± 60)*	490 *(± 20)*	1100 *(± 100)*
C2	500 *(± 1)*	750 *(± 40)*	501 *(± 2)*	1000 *(± 100)*
C3	501 *(± 2)*	790 *(± 70)*	500 *(± 2)*	790 *(± 50)*
C4	499 *(± 2)*	760 *(± 70)*	501 *(± 1)*	700 *(± 100)*
C5	501 *(± 3)*	800 *(± 200)*	500 *(± 2)*	800 *(± 200)*

F

Supportive Information for Chapter 9 – Management under invasion pressure

F.1 Simulated biomass increase

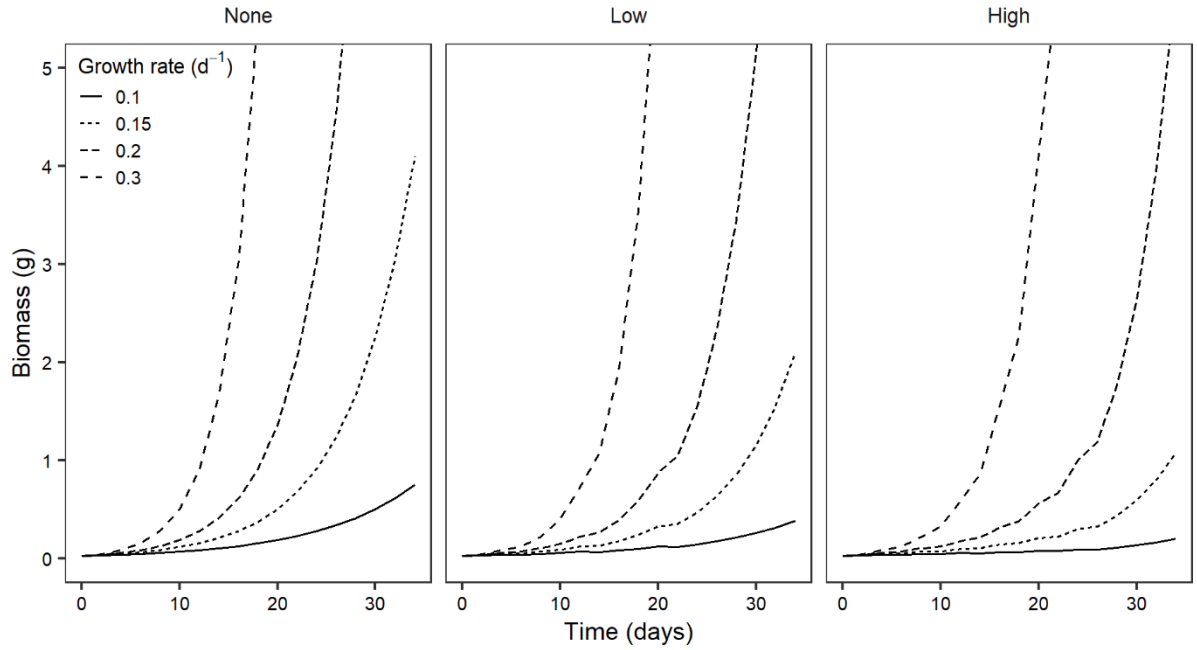

Figure F.1: Simulations of temporal biomass increase under harvesting pressure. *The effect of four relative growth rates on biomass production is depicted and shows a clear difference in produced biomass. The considered removal scenarios include (i) no removal (left), (ii) low-frequency removal (middle) and (iii) high-frequency removal (right). Biomass production of the primary species is assumed to be unaffected by introduction of a secondary species (see Equation 9.1). The actual growth rate is expected to be situated within the range tested here.*

F.2 Experimental results

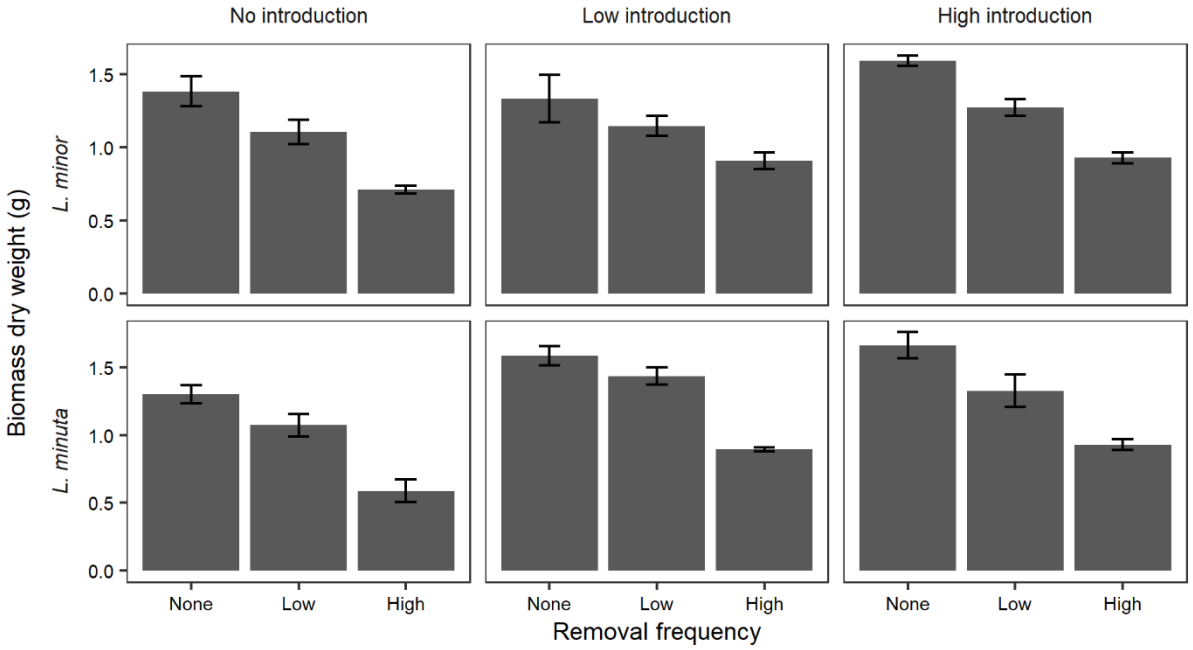

Figure F.2: Total biomass produced in each scenario. *Vertical bars indicate the combined biomass of* Lemna minor *and* L. minuta*, measured over three replicates. Error bars indicate the standard deviation.*

Table F.1: Overview of* p*-values obtained via the two-sample* t*-test for comparing biomass of* Lemna minor *under different management scenarios, with* L. minor *as primary species.

	1	2	3	4	5	6	7	8	9
1	-	0.0469	0.0238	0.0421	NA	NA	0.2835	NA	NA
2	0.0469	-	0.0347	NA	0.0410	NA	NA	0.0537	NA
3	0.0238	0.0347	-	NA	NA	0.4255	NA	NA	0.0964
4	0.0421	NA	NA	-	0.5230	0.2047	0.0798	NA	NA
5	NA	0.0410	NA	0.5230	-	0.0421	NA	0.3974	NA
6	NA	NA	0.4255	0.2047	0.0421	-	NA	NA	0.2434
7	0.2835	NA	NA	0.0798	NA	NA	-	0.0034	0.0007
8	NA	0.0537	NA	NA	0.3974	NA	0.0034	-	0.0143
9	NA	NA	0.0964	NA	NA	0.2434	0.0007	0.0143	-

Table F.2: Overview of* p*-values obtained via the two-sample* t*-test for comparing biomass of* Lemna minuta *under different management scenarios, with* L. minor *as primary species.

	4	5	6	7	8	9
4	-	0.0592	0.0059	0.0308	NA	NA
5	0.0592	-	0.0718	NA	0.0718	NA
6	0.0059	0.0718	-	NA	NA	0.0375
7	0.0308	NA	NA	-	0.0375	0.1357
8	NA	0.0718	NA	0.0375	-	0.0375
9	NA	NA	0.0375	0.1357	0.0375	-

Table F.3: Overview of* p*-values obtained via the two-sample* t*-test for comparing biomass of* Lemna minuta *under different management scenarios, with* L. minuta *as primary species.

	1	2	3	4	5	6	7	8	9
1	-	0.0440	0.0075	0.9926	NA	NA	0.8286	NA	NA
2	0.0440	-	0.0124	NA	0.1836	NA	NA	0.0467	NA
3	0.0075	0.0124	-	NA	NA	0.1757	NA	NA	0.3216
4	0.9926	NA	NA	-	0.2121	0.0185	0.8286	NA	NA
5	NA	0.1836	NA	0.2121	-	0.0132	NA	0.0124	NA
6	NA	NA	0.1757	0.0185	0.0132	-	NA	NA	0.0968
7	0.8286	NA	NA	0.8286	NA	NA	-	0.0168	0.0168
8	NA	0.0467	NA	NA	0.0124	NA	0.0168	-	0.0243
9	NA	NA	0.3216	NA	NA	0.0968	0.0168	0.0243	-

Table F.4: Overview of* p*-values obtained via the two-sample* t*-test for comparing biomass of* Lemna minor *under different management scenarios, with* L. minuta *as primary species.

	4	5	6	7	8	9
4	-	0.2216	0.0594	0.1360	NA	NA
5	0.2216	-	0.0886	NA	0.0594	NA
6	0.0594	0.0886	-	NA	NA	0.0886
7	0.1360	NA	NA	-	0.0888	0.1187
8	NA	0.0594	NA	0.0888	-	0.0594
9	NA	NA	0.0886	0.1187	0.0594	-

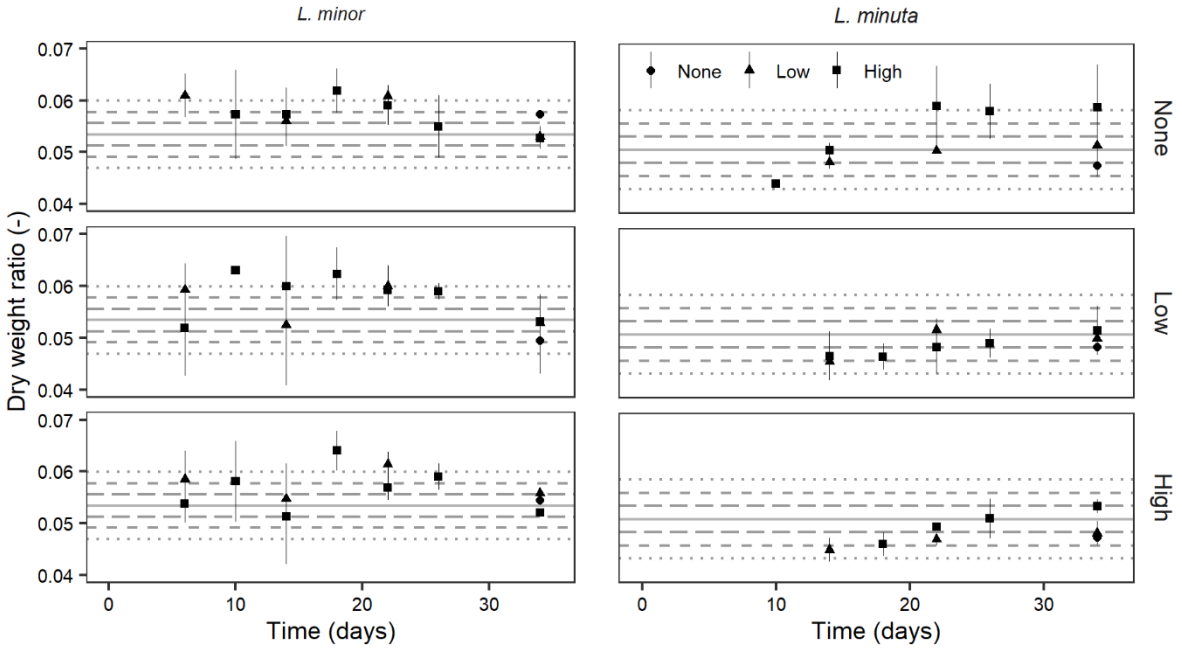

***Figure F.3: Temporal variation in dry weight ratio for* Lemna minor *and* Lemna minuta.** *The solid horizontal line represents the average dry weight ratio at day 34, with longdashed, dashed and dotted lines representing the range including 1, 2 or 3 times the standard deviation, respectively. Dry weight ratios tend to be higher for* L. minor *than for* L. minuta.

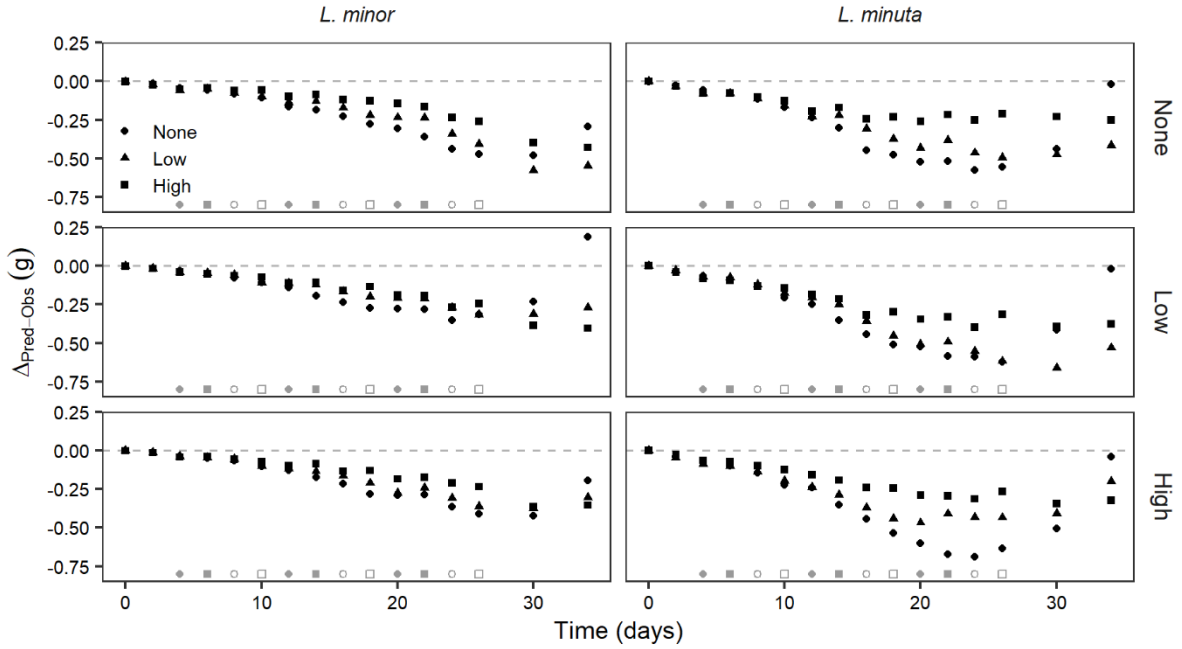

Figure F.4: Error in dry weight between predictions and observations. *All predictions underestimated the obtained biomass, except for the last observations of* L. minor *devoid of biomass removal and under low introduction pressure. The temporal evolution indicates an underestimation of the growth rate, causing errors to keep increasing until the actual growth rate becomes lower than the time-independent fixed rate. Grey symbols represent introduction (circles) and removal (squares) events, with filled symbols indicating the low frequency pressure.*

F.3 Generalised linear mixed effects models

The influence of each fixed effect (i.e. time, removal frequency and introduction pressure) on the obtained biomass was inferred from linear mixed effects models. Biomass was cube root transformed to represent a more symmetrical distribution (see Figure F.5) and all interactions among the fixed effects were considered within the saturated model. As biomass was registered for periods without any treatment (i.e. undisturbed growth from day 0 to day 4 and from day 26 to day 34), the fixed effect 'Time' was divided into three dummy scores, splitting at day 4 (start of the treatment) and day 26 (end of the treatment). Lastly, individual aquarium codes were considered as random effects within the repeated measurement scenario. Model development followed the procedure as explained by Zuur *et al.* (2009). In short, the procedure defined (1) the added value of using mixed effects over ordinary linear models, (2) the random structure (with restricted maximum likelihood (REML) fitting), (3) the fixed structure (with maximum likelihood (ML) fitting and manual backward term selection) and (4) final model fit (with REML) along with assessment of model residuals. Results for both *Lemna minor* and *L. minuta* are presented in the following sections.

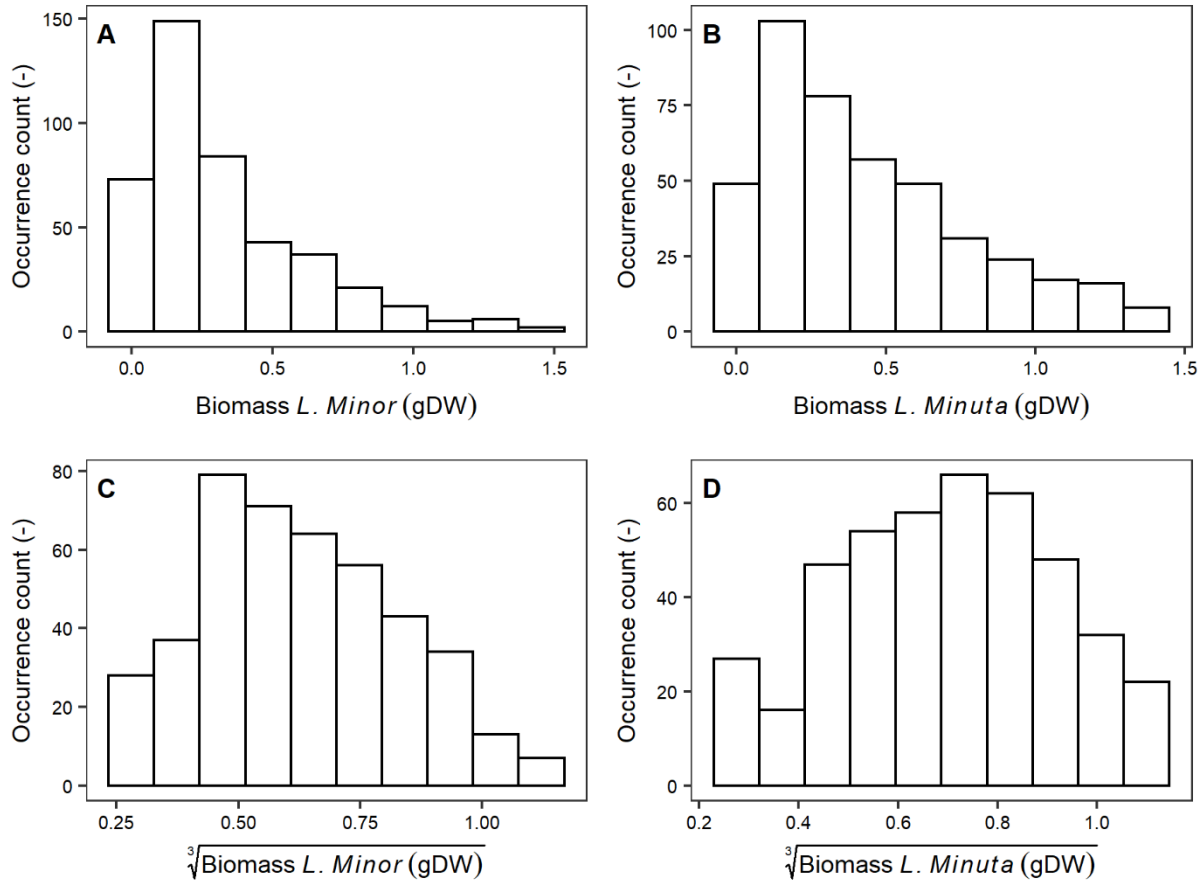

Figure F.5: Distribution of biomass values. *Biomass scores for both* Lemna *spp. showed to be skewed (A:* L. minor*; B:* L. minuta*), while more symmetrical distributions were obtained after cube root transformation (C:* L. minor*; D:* L. minuta*).*

F.3.1 Results for *Lemna minor*

Model selection showed a significantly better fit of the saturated linear mixed effects model over the ordinary linear model (L = 153.8, df = 2, p < 0.001). Assessment of the variance structure illustrated no significant improvements in the Akaike Information Criterion (AIC) by considering a random slope structure for time rather than a random intercept structure, hence no random slopes were included. Lastly, interactions of treatment with the first time period were excluded and showed to improve AIC scores by reducing model complexity (i.e. -2033 versus -2019). No further reductions in model complexity could be performed without causing an increase in AIC scores. The resulting coefficient estimates of the fixed effects are summarised in Table F.5.

***Table F.5: Estimates of the fixed effects coefficients within the linear mixed effects model for* Lemna minor.** *Aside from the estimate, the standard error, degrees of freedom (DF),* t-*value,* p-*value and the range (Lower and Upper) are provided (not reflecting standard confidence intervals).*

Parameter	Estimate	Error	DF	*t*-value	*p*-value	Lower	Upper
β_{o}	0.2976	0.0128	386	23.28	0.000	0.2724	0.3227
β_{T1}	0.0307	0.0011	386	26.94	0.000	0.0284	0.0329
β_{T2}	0.0133	0.0005	386	27.90	0.000	0.0124	0.0143
β_{T3}	0.0198	0.0017	386	11.69	0.000	0.0165	0.0231
β_{OutLow}	-0.0051	0.0175	18	-0.29	0.773	-0.0419	0.0316
$\beta_{OutNone}$	-0.0031	0.0175	18	-0.18	0.862	-0.0398	0.0337
β_{InLow}	0.0037	0.0175	18	0.21	0.834	-0.0330	0.0405
β_{InNone}	0.0003	0.0175	18	0.02	0.987	-0.0365	0.0370
$\beta_{T2:OutLow}$	0.0057	0.0007	386	8.58	0.000	0.0044	0.0070
$\beta_{T2:OutNone}$	0.0109	0.0007	386	16.34	0.000	0.0096	0.0122
$\beta_{T3:OutLow}$	-0.0047	0.0024	386	-1.98	0.049	-0.0094	0.0000
$\beta_{T3:OutNone}$	-0.0018	0.0024	386	-0.74	0.457	-0.0065	0.0029
$\beta_{T2:InLow}$	0.0006	0.0007	386	0.94	0.350	-0.0007	0.0019
$\beta_{T2:InNone}$	-0.0002	0.0007	386	-0.30	0.763	-0.0015	0.0011
$\beta_{T3:InLow}$	0.0002	0.0024	386	0.09	0.926	-0.0045	0.0049
$\beta_{T3:InNone}$	0.0044	0.0024	386	1.84	0.066	-0.0003	0.0091
$\beta_{OutLow:InLow}$	0.0040	0.0247	18	0.16	0.873	-0.0480	0.0560
$\beta_{OutNone:InLow}$	0.0030	0.0247	18	0.12	0.905	-0.0490	0.0550
$\beta_{OutLow:InNone}$	0.0121	0.0247	18	0.49	0.631	-0.0399	0.0641
$\beta_{OutNone:InNone}$	0.0057	0.0247	18	0.23	0.821	-0.0463	0.0577
$\beta_{T2:OutLow:InLow}$	-0.0022	0.0009	386	-2.33	0.020	-0.0040	-0.0003
$\beta_{T2:OutNone:InLow}$	-0.0014	0.0009	386	-1.52	0.130	-0.0033	0.0004
$\beta_{T2:OutLow:InNone}$	-0.0002	0.0009	386	-0.20	0.841	-0.0020	0.0017
$\beta_{T2:OutNone:InNone}$	0.0009	0.0009	386	0.95	0.344	-0.0010	0.0027
$\beta_{T3:OutLow:InLow}$	0.0011	0.0034	386	0.33	0.738	-0.0055	0.0078
$\beta_{T3:OutNone:InLow}$	-0.0139	0.0034	386	-4.09	0.000	-0.0205	-0.0072
$\beta_{T3:OutLow:InNone}$	0.0065	0.0034	386	1.91	0.056	-0.0002	0.0131
$\beta_{T3:OutNone:InNone}$	-0.0040	0.0034	386	-1.19	0.233	-0.0107	0.0026

The obtained model showed to encapsulate most of the variance included within the fixed effects (Figure F.6A) and produced an acceptable quantile-quantile plot (Figure F.6B). Moreover, only limited patterns related to the main effects remained unexplained by the model and were mostly linked with the Time effect (see Figure F.7C), while residuals were nicely distributed around zero for both removal frequency (Figure F.7B) and introduction pressure (Figure F.7A).

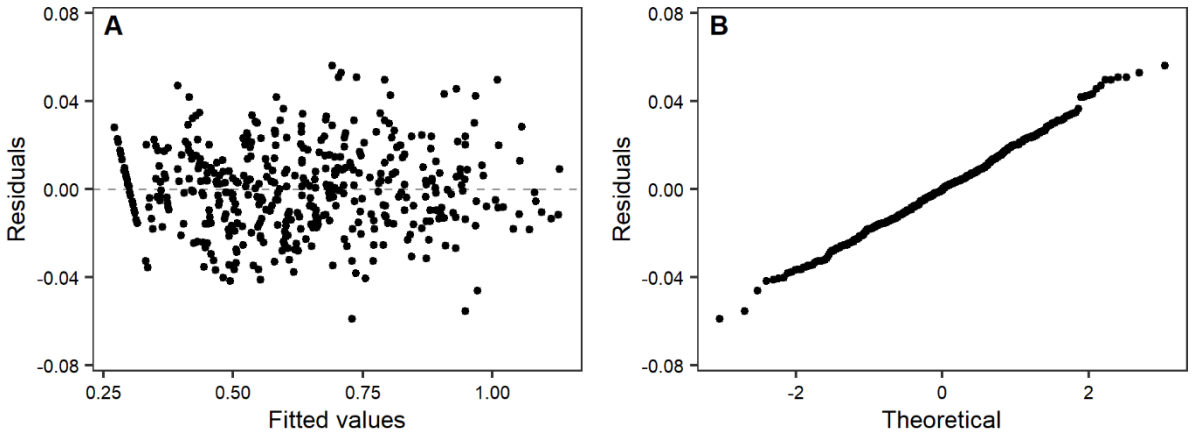

Figure F.6: Residuals of the final linear mixed effects model. *A: Residuals are clearly scattered around zero; B: Quantile-quantile plot supporting acceptable model fit.*

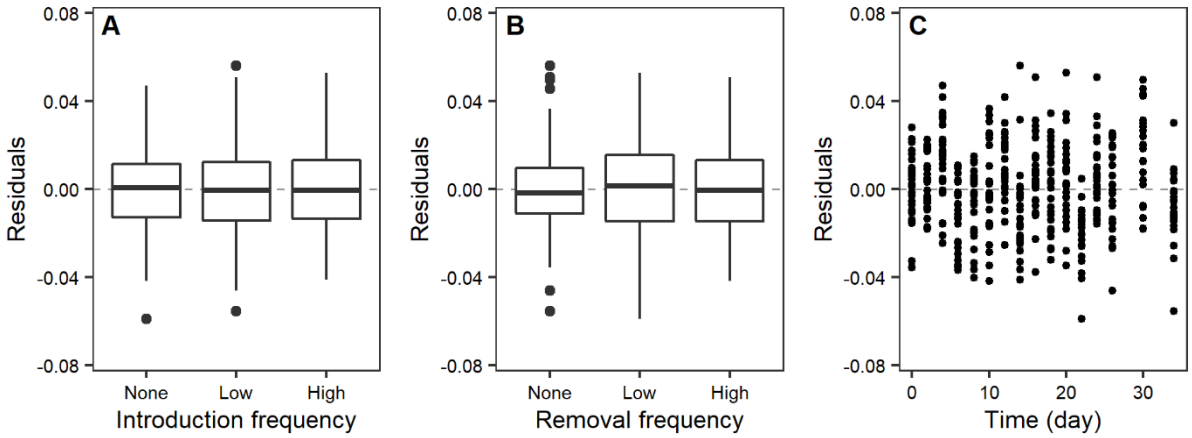

Figure F.7: Effect-specific distribution of model residuals. *A: Distribution of the residuals conditional to the applied introduction frequency; B: Distribution of the residuals conditional to the applied biomass removal frequency; C: Distribution of the residuals conditional to the measurement day. A minor pattern in residual distribution can be observed for the main effect of time, oscillating around zero.*

Finally, the developed model was applied to the original data to visually assess model fit when contrasting observations with predictions. In general, a high model fit is observed for the final model, with predictions clearly following the observed temporal pattern, conditional to the applied treatment (Figure F.8).

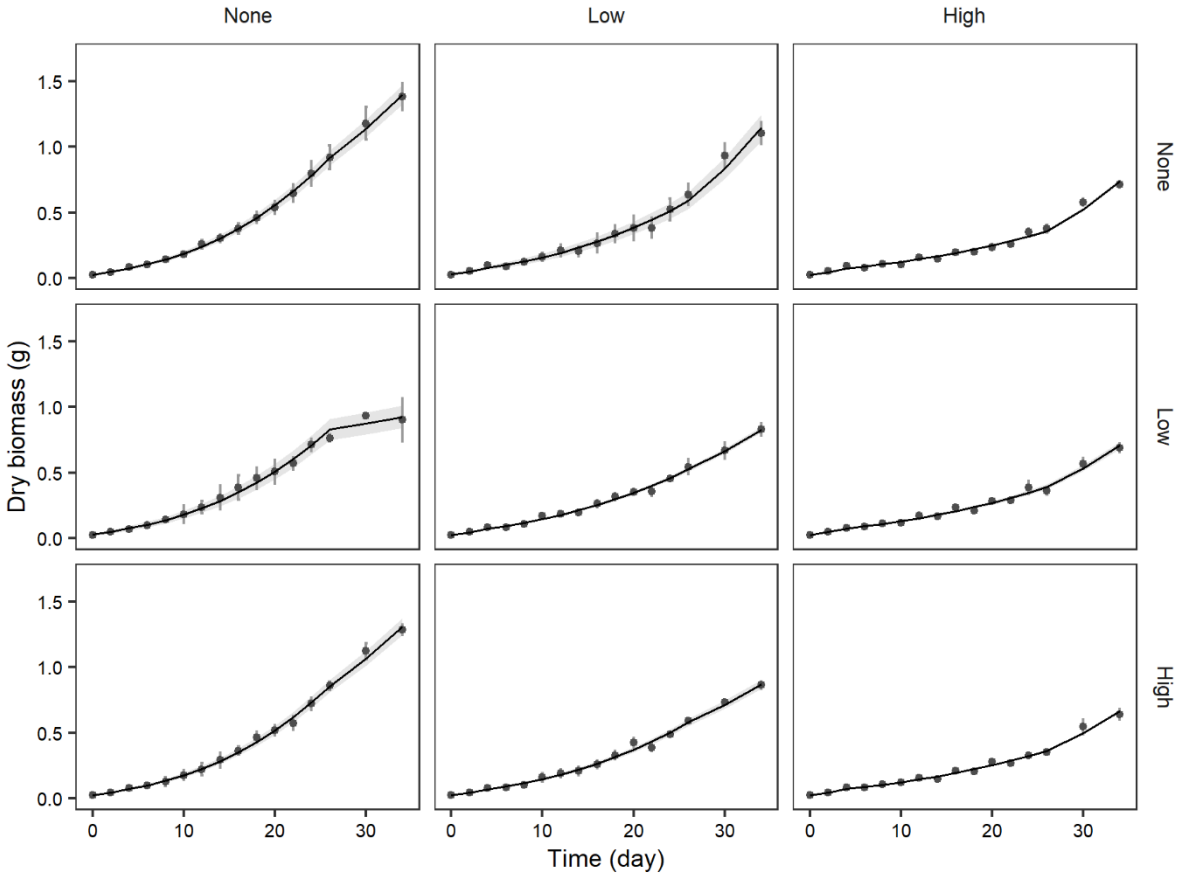

Figure F.8: Model predictions versus observations. *Predictions from the developed linear mixed effects model (black lines) clearly followed the observed temporal patterns (dark grey circles). Observations combined three replicates (vertical error bars indicating the standard deviation), which resulted in three separate predictions, causing a ribbon (light grey zone indicating the standard deviation) to be depicted around the line connecting the mean of all predictions. Rows indicate the introduction pressure, while columns entail the different biomass removal frequencies.*

F.3.2 Results for *Lemna minuta*

Model selection showed a significantly better fit of the saturated linear mixed effects model over the ordinary linear model (L = 177.2, df = 2, p < 0.001). Assessment of the variance structure illustrated no significant improvements in the Akaike Information Criterion (AIC) by considering a random slope structure for time rather than a random intercept structure, hence no random slopes were included. Lastly, interactions of treatment with the first time period were excluded and showed to improve AIC scores by reducing model complexity (i.e. -1922 versus -1910). Further reductions in model complexity and AIC scores were obtained by excluding interactions between the third time interval and introduction pressure. The resulting coefficient estimates of the fixed effects are summarised in Table F.6.

***Table F.6: Estimates of the fixed effects coefficients within the linear mixed effects model for* Lemna minuta.** *Aside from the estimate, the standard error, degrees of freedom (DF),* t-*value,* p-*value and the range (Lower and Upper) are provided (not reflecting standard confidence intervals).*

Parameter	Estimate	Error	DF	*t*-value	*p*-value	Lower	Upper
β_0	0.2948	0.0146	392	20.25	0.0000	0.2662	0.3235
β_{T1}	0.0487	0.0016	392	30.00	0.0000	0.0455	0.0519
β_{T2}	0.0129	0.0010	392	12.90	0.0000	0.0109	0.0148
β_{OutLow}	0.0253	0.0198	18	1.28	0.2184	-0.0164	0.0670
$\beta_{OutNone}$	0.0101	0.0198	18	0.51	0.6155	-0.0316	0.0518
β_{InLow}	0.0185	0.0198	18	0.93	0.3634	-0.0232	0.0601
β_{InNone}	0.0123	0.0198	18	0.62	0.5441	-0.0294	0.0539
β_{T3}	0.0141	0.0014	392	10.01	0.0000	0.0113	0.0168
$\beta_{T2:OutLow}$	0.0048	0.0014	392	3.40	0.0007	0.0020	0.0075
$\beta_{T2:OutNone}$	0.0133	0.0014	392	9.49	0.0000	0.0105	0.0160
$\beta_{T2:InLow}$	0.0005	0.0014	392	0.40	0.6912	-0.0021	0.0032
$\beta_{T2:InNone}$	-0.0024	0.0014	392	-1.75	0.0803	-0.0050	0.0003
$\beta_{OutLow:InLow}$	-0.0508	0.0280	18	-1.81	0.0864	-0.1097	0.0080
$\beta_{OutNone:InLow}$	-0.0240	0.0280	18	-0.86	0.4026	-0.0829	0.0349
$\beta_{OutLow:InNone}$	-0.0394	0.0280	18	-1.41	0.1769	-0.0983	0.0195
$\beta_{OutNone:InNone}$	-0.0307	0.0280	18	-1.09	0.2885	-0.0895	0.0282
$\beta_{T3:OutLow}$	-0.0037	0.0020	392	-1.87	0.0628	-0.0076	0.0002
$\beta_{T3:OutNone}$	-0.0087	0.0020	392	-4.36	0.0000	-0.0126	-0.0047
$\beta_{T2:OutLow:InLow}$	0.0042	0.0019	392	2.19	0.0288	0.0004	0.0080
$\beta_{T2:OutNone:InLow}$	-0.0010	0.0019	392	-0.50	0.6167	-0.0047	0.0028
$\beta_{T2:OutLow:InNone}$	0.0047	0.0019	392	2.48	0.0137	0.0010	0.0085
$\beta_{T2:OutNone:InNone}$	0.0022	0.0019	392	1.15	0.2492	-0.0016	0.0060

The obtained model showed to encapsulate most of the variance included within the fixed effects (Figure F.9A) and produced an acceptable quantile-quantile plot (Figure F.9B). Moreover, only limited patterns related to the main effects remained unexplained by the model and were mostly linked with the Time effect (see Figure F.10C), while residuals were nicely distributed around zero for both removal frequency (Figure F.10B) and introduction pressure (Figure F.10A).

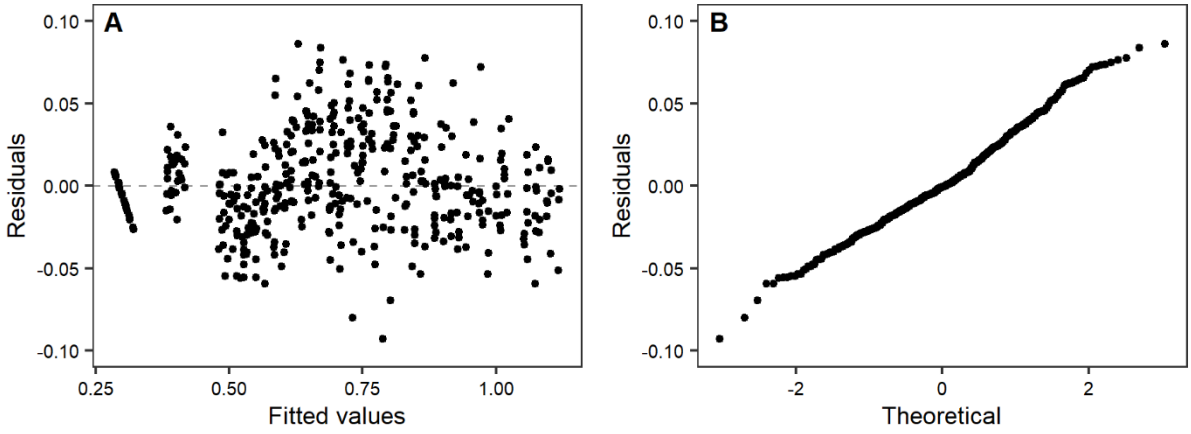

Figure F.9: Residuals of the final linear mixed effects model. *A: Residuals are clearly scattered around zero; B: Quantile-quantile plot supporting acceptable model fit.*

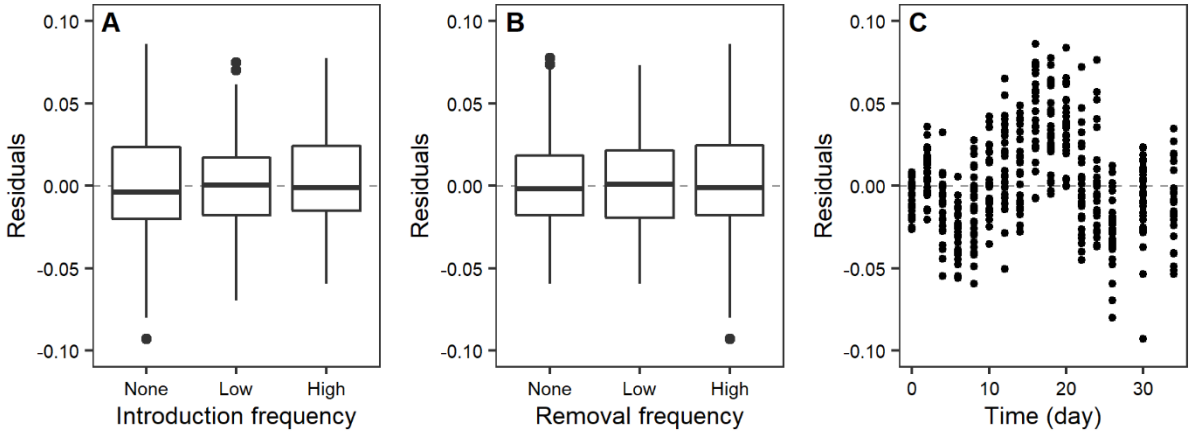

Figure F.10: Effect-specific distribution of model residuals. *A: Distribution of the residuals conditional to the applied introduction frequency; B: Distribution of the residuals conditional to the applied biomass removal frequency; C: Distribution of the residuals conditional to the measurement day. A relatively clear pattern in residual distribution can still be observed for the main effect of time, oscillating around zero and suggesting that other transformations or link-functions require consideration.*

Finally, the developed model was applied to the original data to visually assess model fit when contrasting observations with predictions. In general, a good model fit is observed for the final model, with predictions clearly following the observed temporal pattern, conditional to the applied treatment (Figure F.11). Nevertheless, a higher discrepancy can be observed for *Lemna minuta* compared to *L. minor* (see Figure F.8).

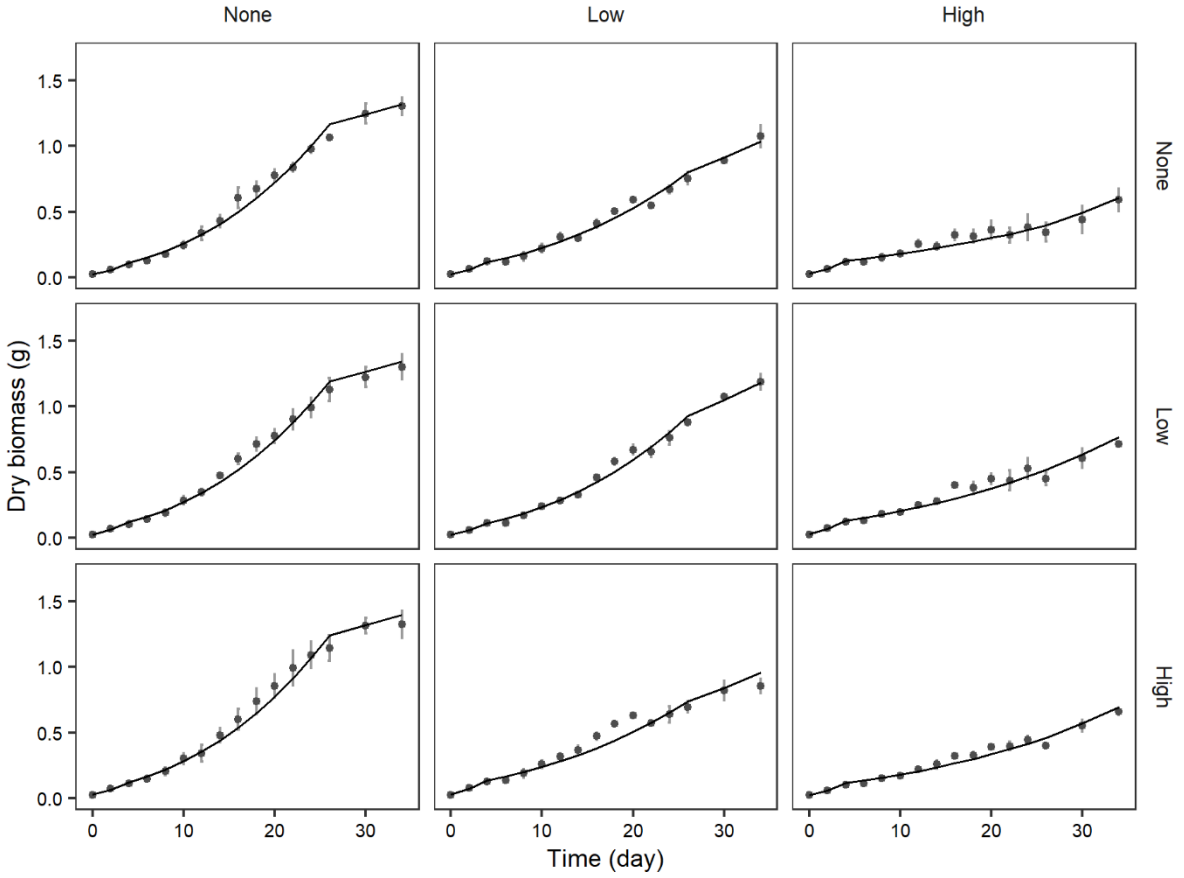

Figure F.11: Model predictions versus observations. *Predictions from the developed linear mixed effects model (black lines) clearly followed the observed temporal patterns (dark grey circles). Observations combined three replicates (vertical error bars indicating the standard deviation), which resulted in three separate predictions, causing a ribbon (light grey zone indicating the standard deviation) to be depicted around the line connecting the mean of all predictions. Rows indicate the introduction pressure, while columns entail the different biomass removal frequencies.*

Curriculum vitae

Personal information

Name	Van Echelpoel
First name	Wout
Day of birth	21 June 1990
Nationality	Belgian
Contact	wout.vanechelpoel@gmail.com
LinkedIn	linkedin.com/in/wout-van-echelpoel-a8456641/
ORCID	0000-0001-9636-5861

Education

Full-time studies

04/2014 – Today	Doctoral Schools of Bioscience Engineering, Ghent University
09/2013 – 07/2014	Advanced Master of Science in Technology for Integrated Water Management (greatest distinction), Ghent University and University of Antwerp **Thesis**: Microplastics in a biological wastewater treatment plant and the surrounding freshwater environment in Flanders: quantitative assessment
09/2011 – 07/2013	Master of Science in Bioscience Engineering: Environmental Technology, Ghent University (great distinction) **Thesis**: Micro-CT als innovatieve visualisatietechniek van microplastics in mariene organismen
09/2008 – 09/2011	Bachelor of Science in Bioscience Engineering: Environmental Technology, University of Antwerp
09/2002 – 06/2008	Latin-Mathematics, Groenendaal Merksem

Individual courses

05/2019 – 06/2019	Data Mining (ICES Data Analysis), Ghent University
09/2018 – 06/2019	Java I & II, Het Perspectief
02/2018 – 06/2018	Start to program: Java, Het Perspectief
02/2017 – 01/2018	Spanish breakthrough (A2) & waystage (A1), Het Perspectief
02/2015 – 06/2015	Practical English 5, Ghent University
09/2014	Introduction to Computational Fluid Dynamics, Ghent University

Publications

Articles in peer reviewed scientific journals (A1)

Sampantamit, T.; Ho, L.; **Van Echelpoel, W.**; Lachat, C. and Goethals, P. (2020) Links and Trade-Offs between Fisheries and Environmental Protection in Relation to the Sustainable Development Goals in Thailand. Water 12, doi: 10.3390/w12020399.

Deknock, A.; De Troyer, N.; Houbraken, M.; Dominguez-Granda, L.; Nolivos, I.; **Van Echelpoel, W.**; Forio, M. A. E.; Spanoghe, P. and Goethals, P. (2019) Distribution of agricultural pesticides in the freshwater environment of the Guayas river basin (Ecuador). Science of the Total Environment 646, 996-1008, doi: 10.1016/j.scitotenv.2018.07.185.

Ho, L.; Thas, O.; **Van Echelpoel, W.** and Goethals, P. (2019) A Practical Protocol for the Experimental Design of Comparative Studies on Water Treatment. Water 11, doi: 10.3390/w11010162.

Van Echelpoel, W.; Forio, M. A. E.; Van der heyden, C.; Bermúdez, R.; Ho, L.; Rosado Moncayo, A. M.; Parra Narea, R. N.; Dominguez Granda, L. E.; Sanchez, D. and Goethals, P. L. M. (2019) Spatial Characteristics and Temporal Evolution of Chemical and Biological Freshwater Status as Baseline Assessment on the Tropical Island San Cristóbal (Galapagos, Ecuador). Water 11, doi: 10.3390/w11050880.

Ho, L.; Pham, T. D.; **Van Echelpoel, W.**; Muchene, L.; Shkedy, Z.; Alvarado, A.; Espinoza-Palacios, J.; Arevalo-Durazno, M.; Thas, O. and Goethals, P. (2018) A Closer Look on Spatiotemporal Variations of Dissolved Oxygen in Waste Stabilization Ponds Using Mixed Models. Water 10, doi: 10.3390/w10020201.

Ho, L.; Pompeu, C.; **Van Echelpoel, W.**; Thas, O. and Goethals, P. (2018) Model-Based Analysis of Increased Loads on the Performance of Activated Sludge and Waste Stabilization Ponds. Water 10, doi: 10.3390/w10101410.

Ho, L.; **Van Echelpoel, W.**; Charalambous, P.; Gordillo, P. A.; Thas, O. and Goethals, P. (2018) Statistically-Based Comparison of the Removal Efficiencies and Resilience Capacities between Conventional and Natural Wastewater Treatment Systems: A Peak Load Scenario. Water 10, doi: 10.3390/w10030328.

Ho, L. T.; Pham, D. T.; **Van Echelpoel, W.**; Alvarado, A.; Espinoza-Palacios, J. E.; Arevalo-Durazno, M. B. and Goethals, P. L. M. (2018) Exploring the influence of meteorological conditions on the performance of a waste stabilization pond at high altitude with structural equation modeling. Water Science and Technology.

Jerves-Cobo, R.; Córdova-Vela, G.; Iñiguez-Vela, X.; Díaz-Granda, C.; **Van Echelpoel, W.**; Cisneros, F.; Nopens, I. and Goethals, L. P. (2018) Model-Based Analysis of the Potential of Macroinvertebrates as Indicators for Microbial Pathogens in Rivers. Water 10, doi: 10.3390/w10040375.

Van Echelpoel, W.; Forio, M. A. E.; Van Butsel, J.; Lock, K.; Utreras, J. A. D.; Dominguez-Granda, L. E. and Goethals, P. L. M. (2018) Macroinvertebrate functional feeding group structure along an impacted tropical river: The Portoviejo River (Ecuador). Limnologica 73, 12-19, doi: 10.1016/j.limno.2018.10.001.

Van Echelpoel, W. and Goethals, P. L. M. (2018) Variable importance for sustaining macrophyte presence via random forests: data imputation and model settings. Scientific Reports 8, 14557, doi: 10.1038/s41598-018-32966-2.

Ho, L. T.; **Van Echelpoel, W.** and Goethals, P. L. M. (2017) Design of waste stabilization pond systems: A review. Water Research 123, 236-248, doi: 10.1016/j.watres.2017.06.071.

Nguyen, T. H. T.; Boets, P.; Lock, K.; Forio, M. A. E.; **Van Echelpoel, W.**; Van Butsel, J.; Utreras, J. A. D.; Everaert, G.; Granda, L. E. D.; Hoang, T. H. T. and Goethals, P. L. M. (2017) Water quality related macroinvertebrate community responses to environmental gradients in the Portoviejo River (Ecuador). Annales de Limnologie - International Journal of Limnology 53, 203-219, doi: 10.1051/limn/2017007.

Forio, M. A. E.; **Van Echelpoel, W.**; Dominguez-Granda, L.; Mereta, S. T.; Ambelu, A.; Hoang, T. H.; Boets, P. and Goethals, P. L. M. (2016) Analysing the effects of water quality on the occurrence of freshwater macroinvertebrate taxa among tropical river basins from different continents. AI Communications, 1-21.

Gobeyn, S.; Bennetsen, E.; **Van Echelpoel, W.**; Everaert, G. and Goethals, P. L. M. (2016) Impact of abundance data errors on the uncertainty of an ecological water quality assessment index. Ecological Indicators 60, 746-753, doi: 10.1016/j.ecolind.2015.07.031.

Van Echelpoel, W.; Boets, P. and Goethals, P. L. M. (2016) Functional Response (FR) and Relative Growth Rate (RGR) Do Not Show the Known Invasiveness of Lemna minuta (Kunth). PLoS ONE 11, e0166132, doi: 10.1371/journal.pone.0166132.

Articles in peer reviewed scientific journals (A2)

Arévalo, M. B.; **Van Echelpoel, W.**; Alvarado, A. O.; Goethals, P. L. M. and Larriva, J. B. (2017) Análisis espacial-temporal de procesos relacionados con concentraciones de oxígeno disuelto en lagunas de maduración. MASKANA 8, 115-123, doi: 10.18537/mskn.08.02.09.

Articles in conference proceedings (P1)

Van Echelpoel, W.; Boets, P.; Landuyt, D.; Gobeyn, S.; Everaert, G.; Bennetsen, E.; Mouton, A. and Goethals, P. L. M. (2015) Species distribution models for sustainable ecosystem management in *Developments in Environmental Modelling* Vol. 27 (eds Y.-S. Park, S. Lek, C. Baehr and S. E. Jørgensen) Ch. 6, 115-134 (Elsevier, The Netherlands).

In preparation

Bruneel, S.; Ho, L.; **Van Echelpoel, W.**; Schoeters, A.; Raat, H.; Moens, T.; Bermudez, R.; Luca, S. and Goethals, P. (under review) Assessment of video transect designs for reef fish assemblages

Forio, M. A. E.,; Villa-Cox, G.; **Van Echelpoel, W.**; Ryckebusch, H.; Lock, K.; Spanoghe, P.; Deknock, A.; De Troyer, N.; Nolivos, I.; Dominguez, L.; Speelman, S. and Goethals, P. (under review) Bayesian Belief Network models as trade-off tools of ecosystem services in the Guayas River Basin in Ecuador

Bruneel, S.; **Van Echelpoel, W.**; Ho, L.; Raat, H.; Schoeters, A.; De Troyer, N.; Sor, R.; Ponton Cevallos, J.; Vandeputte, R.; Van der heyden, C.; De Saeyer, N.; Forio, M. A. E.; Bermúdez, J. R.; Dominguez, L.; Luca, S.; Moens, T. and Goethals, P. (submitted) Assessing the drivers behind the structure and diversity of fish assemblages associated with rocky shores in the Galapagos archipelago

Van Echelpoel, W.; Bruneel, S. and Goethals, P. L. M. (submitted) Empirical evaluation of four data imputation methods for incomplete environmental data with varying levels of available information

Abera, B.; **Van Echelpoel, W.**; Tytgat, B.; Kibret, M.; Spanoghe, P.; Ayalew, D.; Adgo, E.; Nyssen, J.; Goethals, P. L. M. and Verleyen, E. (in preparation) Pesticide residues and their effect on macroinvertebrate community composition in the Lake Tana Basin, Ethiopia

Bruneel, S.; Ponton-Cevallos, J.; Riascos, L.; **Van Echelpoel, W.**; Bermudez, R. and Goethals, P. (in preparation) Fine-scale spatial and temporal dynamics of subtidal reef fish assemblages in the Galapagos

Van Echelpoel, W.; Bruneel, S. and Goethals, P. L. M. (in preparation) Speed-performance trade-off in threshold selection during data pre-processing to support water management

Van Echelpoel, W.; Forio, M. A. E. and Goethals, P. L. M. (in preparation) Abiotic habitat suitability models as first-level assessment for restoration potential and invasion vulnerability

Conference contributions

Oral presentations at national and international conferences and symposia

Deknock, A.*; De Troyer, N.; Houbraken, M.; Dominguez-Granda, L.; Nolivos, I.; **Van Echelpoel, W.**; Forio, M. A. E.; Spanoghe, P. and Goethals, P. (2019) Distribution of agricultural pesticides in the freshwater environment of the Guayas river basin (Ecuador). IWA-IDB Innovation Conference on Sustainable Water Use in Cities, Industry and Agriculture, Guayaquil, Ecuador

Van Echelpoel, W.* and Goethals, P. L. M. (2019) Forecasting the Effects of Elevated Nutrient Levels on Vulnerability to Invasion – Traits versus Observations. IWA-IDB Innovation Conference on Sustainable Water Use in Cities, Industry and Agriculture, Guayaquil, Ecuador

Van Echelpoel, W.* and Goethals, P. L. M. (2019) Powerful Plants for Watery Wetlands – Key Issues for Implementing Artificial Multi-purpose Wetlands. IWA-IDB Innovation Conference on Sustainable Water Use in Cities, Industry and Agriculture, Guayaquil, Ecuador

Villa-Cox, G.*; Forio, M. A. E.; **Van Echelpoel, W.**; Ryckebusch, H.; Lock, K.; Spanoghe, P.; Deknock, A.; De Troyer, N.; Nolivos, I.; Dominguez, L.; Speelman, S. and Goethals, P. L. M. (2019) Bayesian Belief Network model as a tradeoff tool to estimate ecosystem services: case study of the Guayas River Basin, Ecuador. IWA-IDB Innovation Conference on Sustainable Water Use in Cities, Industry and Agriculture, Guayaquil, Ecuador

Forio, M. A. E.*; Ryckebusch, H.; **Van Echelpoel, W.**; Villa Cox, G. and Goethals, P. (2018) BBN models as trade-off tools for ecosystem services. 10th International Conference on Ecological Informatics, Jena, Germany

Forio, M. A. E.; **Van Echelpoel, W.**; De Troyer, N.; Deknock, A.; Dominguez, L. and Goethals, P.* (2018) Application of BBN-models to link aquatic invertebrate traits to environmental river conditions in the Guayas basin (Ecuador). 10th International Conference on Ecological Informatics, Jena, Germany

Jerves Cobo, R.*; Cordova Vela, G.; Iñiguez Vela, X.; Díaz Granda, C.; **Van Echelpoel, W.**; Cisneros, F.; Nopens, I. and Goethals P. (2018) Model-based analysis of the potential of macroinvertebrates as indicators for microbial pathogens in rivers. NAEM 2018 Netherlands annual ecology meeting, Lunteren, The Netherlands

Van Echelpoel, W.*; Forio, M. A. E.; Van der Heyden, C.; Bermúdez, R. J.; Ho, L.; Rosado Moncayo, A. M.; Parra Narea, R. N.; Dominguez Granda, L. E. and Goethals P. L. M. (2018) Temporal change of freshwater quality on a tropical island (San Cristóbal, Galápagos). AQUATROP, Quito, Ecuador

Ho, L. T.*; **Van Echelpoel, W.**; Duy, T. P.; Espinoza-Palacios, J. E.; Arevalo-Durazno, M. B.; Muchene, L.; Shkedy, Z.; Alvarado, A.; Thas, O. and Goethals, P. (2017) Spatiotemporal effects on oxygen level in waste stabilization pond at high altitude. 5th IWA BeNeLux Regional Young Water Professionals Conference, Ghent, Belgium

Jerves Cobo, R.*; Iñiguez Vela, X.; Cordova Vela, G.; Díaz Granda, C.; **Van Echelpoel, W.**; Cisneros Espinoza, F.; Nopens, I. and Goethals, P. (2016) Macroinvertebrate based mathematical models for the prediction of microbial pathogens in rivers. 3rd International conference on Big Data Analysis & Data Mining, London, UK

* = Presenter

Poster presentations at national and international conferences and symposia

Bruneel, S.; Vanden Bulcke, M.; Ponton-Cevallos, J.; Riascos, L.; **Van Echelpoel, W.**; Bermudez, R.; Moens, T. and Goethals, P. (2020) Fine-scale dynamics in reef fish assemblages: Implications for monitoring. 25th National Symposium of Applied Biological Sciences, Liège, Belgium

Van Echelpoel, W.; Forio, M. A. E. and Goethals, P. (2019) Macrophyte-specific habitat suitability scores as first-level assessment of restoration potential and invasion vulnerability. 24th National Symposium of Applied Biological Sciences, Ghent, Belgium

Choeurn, K.; Sor, R.; Tran, T.; Le, D.; Sambo, Y.; Ngo, Q.; Ho, L.; **Van Echelpoel, W.**; Forio, M. A. E.; Truong, T. and Goethals, P. L. M. (2018) Macroinvertebrate diversity assessment in Co-chien river (Ben Tre province, Vietnam). Sustainable Agriculture in Cambodia: Current Knowledge Applications and Future Needs, UNICAM Conference, Siem Reap, Cambodia

Phan, S.; Truong, T.; Serey, M.; Ho, L.; Sor; R.; **Van Echelpoel, W.**; Forio, M. A. E. and Goethals, P. L. M. (2018) Assessment of macroinvertebrate biodiversity in Ham Luong river (Ben Tre province, Vietnam). Sustainable Agriculture in Cambodia: Current Knowledge Applications and Future Needs, UNICAM Conference, Siem Reap, Cambodia

Soum, S.; Truong, T.; Tran, T.; Le, D.; Sor, R.; Ho, L.; **Van Echelpoel, W.**; Forio, M. A. E.; Ngo, Q. and Goethals, P. L. M. (2018) Assessment of water quality in Ben Tre river – Mekong Delta, Vietnam. Sustainable Agriculture in Cambodia: Current Knowledge Applications and Future Needs, UNICAM Conference, Siem Reap, Cambodia

Sros, M.; Truong, T.; Sor, R.; Ho, L.; **Van Echelpoel, W.**; Forio, M. A. E.; Ngo, Q. and Goethals, P. L. M. (2018) Assessment of macroinvertebrates in Ba Lai river (Ben Tre province, Vietnam). Sustainable Agriculture in Cambodia: Current Knowledge Applications and Future Needs, UNICAM Conference, Siem Reap, Cambodia

Ho, L. T.; **Van Echelpoel, W.**; Duy, T. P.; Espinoza-Palacios, J. E.; Arevalo-Durazno, M. B.; Muchene, L.; Shkedy, Z.; Alvarado, A.; Thas, O. and Goethals, P. (2017) Spatiotemporal effects on oxygen level in waste stabilization pond at high altitude. 5th IWA BeNeLux Regional Young Water Professionals Conference, Ghent, Belgium

Ho, L. T.; Duy, T. P.; **Van Echelpoel, W.**; Alvarado, A. and Goethals, P. (2016) Temporal and spatial variations of dissolved oxygen, pH, and chlorophyll a in waste stabilization pond system at high altitude. Belmundo Water Talks, Ghent, Belgium

Scientific consulting

Van Butsel, J.; Donoso, N.; Gobeyn, S.; De Troyer, N.; **Van Echelpoel, W.**; Lock, K.; Bwambale, G.; Muganzi, E.; Muhangi, C.; Nalumansi, N.; Peeters, L. and Goethals, P. L. M. (2017) Ecological water quality assessment of the Mpanga catchment, Western Uganda. Ghent University and Protos, Ghent, Belgium, 37 pp. (contact: Peter L. M. Goethals, Peter.Goethals@UGent.be; 2016-2017)

Educational activities

Tutor of master theses

Baert, K., 2016-2017, The role of Asellus aquaticus on organic matter degradation in constructed wetlands. Ghent University, MSc. Bioscience Engineering. Promotor: Goethals, P. L. M.; Tutor: Donoso, N. and **Van Echelpoel, W.**

Ryckebusch, H., 2016-2017, Ecosystem services analysis in the Guayas river basin in Ecuador. Ghent University, MSc. Bioscience Engineering. Promotor: Goethals, P. L. M.; Tutor: Forio, M. A. E. and **Van Echelpoel, W.**

Schockaert, I., 2016-2017, Environmental sustainability assessment of land use options in the Guayas river basin (Ecuador). Ghent University, MSc. Bioscience Engineering. Promotor: Goethals, P. L. M.; Tutor: Forio, M. A. E. and **Van Echelpoel, W.**

Charalambous, P., 2015-2016, Comparative performance evaluation of pond systems for wastewater treatment. Ghent University, MSc. Environmental Sanitation. Promotor: Goethals, P. L. M.; Tutor: Ho, L. T. and **Van Echelpoel, W.**

Lopez Gordillo, A. P., 2015-2016, Comparative performance evaluation of constructed wetland systems for wastewater treatment. Ghent University, MSc. Environmental Sanitation. Promotor: Goethals, P. L. M.; Tutor: Ho, L. T. and **Van Echelpoel, W.**

Practical exercises at Ghent University

Master of Science in Aquaculture

Aquatic Ecology

Master of Science in Environmental Sanitation

Environmental Ecology
Natural Systems for (Waste)Water Treatment
Water Quality Management
Water Quality Modelling

Master of Science in Bioscience Engineering

Aquatische en Terrestrische Ecologie
Ecotechniek en Natuurbouw
Ecotechnologie
Waterkwaliteitsbeheer

Advanced Master of Science in Environmental Sanitation and Management

Ecology and Environmental Microbiology

Advanced Master of Science in Technology for Integrated Water Management (in cooperation with University of Antwerp)

Ecological Engineering

International training courses

Water Quality Management, CONSEA Project, An Giang University, An Giang, Vietnam. March 2018

Water Quality Modelling, CONSEA Project, An Giang University, An Giang, Vietnam. March 2018

Ecological Engineering, Inter-university Master Program on Engineering Sciences for Water Resources Management, Cuenca University, Cuenca, Ecuador. July 2015

Miscellaneous

Committees

2019

Co-chair of the organising committee for the first IWA-IDB Innovation Conference on Sustainable Water Use in Cities, Industry and Agriculture, Guayaquil, Ecuador (30 September – 03 October 2019) (including planning, website development and maintenance, editorial activity of programme book and practical organisation presentations)

2015 – 2019

Advisory board of inter-university Advanced Master Technology for Integrated Water Management

2014 – 2018
Inter-university education committee Advanced Master Technology for Integrated Water Management
2015 – 2017
Organising committee annual B-IWA Nocturnal
2015
Organising committee starting event Natural Capital

International sampling campaigns

04/2018
Lake Tana, tributaries and wetlands. Bahir Dar, Ethiopia
03/2018 – 04/2018
Mekong river delta. Ben Tre, Vietnam
07/2017 – 08/2017
Coastal water quality. Galapagos Islands, Ecuador
12/2016
Mpanga river catchment. Fort Portal, Uganda
10/2016 – 11/2016
Freshwater quality. Galapagos Islands, Ecuador
07/2016 – 08/2016
Guayas river basin. Guayaquil, Ecuador
07/2015 – 08/2015
Portoviejo river basin. Portoviejo, Ecuador

Reviews

2020
Water (1)
2019
Heliyon (1); *Water* (1)
2018
Biological Invasions (1)
2017
Biological Invasions (1)

Scholarships & grants

2019
Personal grant for conference participation from the Flemish Fund of Scientific Research (FWO)
2018
Personal grant for conference participation from the Commission Scientific Research from the Faculty of Bioscience Engineering (CWO)
2017
Personal grant for short research stay from the Flemish Fund of Scientific Research (FWO)