Ghent University
Faculty of Bioscience Engineering

DEVELOPMENT AND APPLICATION OF A FRAMEWORK FOR MODEL STRUCTURE EVALUATION IN ENVIRONMENTAL MODELLING

ir. Stijn Van Hoey

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (Ph.D) in Applied Biological Sciences

Academic year 2015-2016

Members of the Jury:

Prof. dr. ir. Korneel Rabaey
Laboratory of Microbial Ecology and Technology (LabMET)
Ghent University, Belgium

Prof. dr. ir. Niko Verhoest
Laboratory of Hydrology and Water Management
Ghent University, Belgium

Prof. dr. ir. Wim Cornelis
Department of Soil Management
Ghent University, Belgium

Prof. dr. ir. Peter Vanrolleghem
ModelEAU
Université Laval, Quebec, Canada

dr. Hans van der Kwast
UNESCO-IHE
Delft, The Netherlands

dr. Jiri Nossent
Flanders Hydraulics Research
Antwerp, Belgium
Vakgroep Hydrologie en Waterbouwkunde
Vrije Universiteit Brussel, Belgium

Supervisors:

Prof. dr. ir. Ingmar Nopens
Department of Mathematical Modelling,
Statistics and Bioinformatics (BIOMATH)
Ghent University, Belgium

Prof. dr. ir. Piet Seuntjens
Flemish Institute of Technological Research
VITO, Belgium
Department of Soil Management
Ghent University, Belgium

Dean:

Prof. dr. Marc Van Meirvenne

Rector:

Prof. dr. Anne De Paepe

ir. Stijn Van Hoey

DEVELOPMENT AND APPLICATION OF A FRAMEWORK FOR MODEL STRUCTURE EVALUATION IN ENVIRONMENTAL MODELLING

Thesis submitted in fulfillment of the requirements for the degree of

Doctor (Ph.D) in Applied Biological Sciences

Academic year 2015-2016

Dutch translation of the title:
Opbouw en toepassing van raamwerk voor modelstructuur evaluatie van modellen voor milieutoepassingen

Please refer to this work as follows:
Stijn Van Hoey (**2016**). *Development and application of a framework for model structure evaluation in environmental modelling*, PhD Thesis, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium.

ISBN 978-90-5989-906-3

Dankwoord

Tijdens dit doctoraat heb ik een atypisch, maar zeer boeiend traject afgelegd. Dit boekje is het academische orgelpunt, maar de visie en kennis die ik opbouwde zal ik steeds met me meedragen. Ik heb de voorbije jaren het geluk gehad veel boeiende mensen, die op allerhande manieren hebben bijgedragen aan dit werk, te leren kennen. Ik wil hen dan ook graag bedanken.

Financieel werd het doctoraat mogelijk gemaakt door de steun van het Vlaams Instituut voor Technologisch Onderzoek (VITO), waarvoor ik oprecht dankbaar ben. Daarnaast heb ik de kans gekregen om de laatste jaren als assistent aan de universiteit Gent te kunnen werken. Ik dank dan ook de professoren van de vakgroep om me die kans te geven. Ik heb het assistentschap ontzettend graag uitgevoerd.

Een speciale dank aan mijn promotoren. Volgens de huidige wetenschappelijke cultuur ben ik allerminst een groots academicus. Ik heb gelukkig het voorrecht gehad dat zij meer in rekening brachten dan enkel het aantal A1-vermeldingen. Voor het blijvende vertrouwen, de continue ondersteuning en de geboden vrijheid ben ik mijn promotoren enorm dankbaar. Ik heb van hen de kans gekregen een wetenschappelijk concours te mogen volbrengen waarbij ik zelf mocht bijleren, de tijd kreeg om op zoek te gaan en mislukkingen mocht maken. Niet enkel datgene uitwerken dat direct (publicatie)succes zou opleveren, maar tijd investeren in de uitwerking van transparant onderzoek en een langetermijnvisie. Wetenschap in dienst van anderen en van het algemeen maatschappelijk nut. Ingmar en Piet, merci!

Ik heb het voorrecht gehad in verschillende omgevingen te mogen werken en veel fijne collega's te leren kennen. De RMA groep van VITO was een aangename omgeving om in te vertoeven, zowel professioneel als tijdens ontspannende activiteiten en losse babbels. Een belangrijk deel van het rekenwerk en de inhoudelijke uitwerking is te danken aan de ondersteuning die ik kreeg in VITO.

Het project in het waterbouwkundig laboratorium heeft inspiratie en richting gegeven aan het doctoraat. Thomas en Fernando, merci voor de gezellige maandagen samen in het Labo. Het project voor VMM toonde aan dat de aanpak en ontwikkelingen voor modelevaluatie effectief bruikbaar zijn. Jef, dankjewel voor je

uitstekende werk bij de uitvoering van de modelevaluatie. De samenwerking met Koen (Antea) omtrent grondwatermodellering, zal ik altijd een voorbeeld vinden van hoe consultancy en academici complementair kunnen zijn. De output van de projecten en samenwerkingen maken slechts beperkt deel uit van de tekst in dit proefschrift, maar ze zijn stuk voor stuk belangrijk geweest om de visie van dit doctoraat mee vorm te geven.

Ik ben echter de laatste jaren het meeste vergroeid geraakt met de vakgroep. Ik had er het geluk in een omgeving te mogen werken vol gedreven onderzoekers en gemotiveerde lesgevers. I sincerely want to thank all the members of the department, especially the Biomath-team. I have always enjoyed the warm atmosphere, friendly encouragements, the dedication and, off course, the many great *social* activities. The friday-drinks, *part of the weekend never dies* weekends, *bierenplezier* boardgame evenings, lots(!) of desserts,... It has been a great time! Special thanks goes to all the people who joined me in the *simulation lab* during all those years, as well as to my partners in crime for teaching several courses with, certainly Elena, Marlies, Chaïm and Wouter. Ik zou ook graag Joris, mijn persoonlijke python-goeroe maar bovenal een fantastisch persoon en Timothy, een grootse grote die zichzelf enorm kan onderschatten, extra willen bedanken voor al de support.

Ik heb ook het geluk gehad zeer fijne thesis- en bachelor studenten te mogen begeleiden. Stijn en Robin, het was me een genoegen jullie te *adopteren*. Jullie hebben straf werk geleverd en zijn gewoon super om mee aan de slag te gaan. De Bogota-case van Elke was een interessant en zeer leerrijk project. Eline, Marjolein, Jasmine en Karen, merci om onze git-proefkonijnen te zijn.

Ik zou ook graag Katrien willen bedanken voor de samenwerking in de eindfase van het doctoraat. Onze discussies en je gedrevenheid werkten heerlijk aanstekelijk. De samenwerking heeft enorm bijgedragen aan de moreel bij de finale afwerking van deze tekst.

Daarnaast een merci aan mijn vrienden buiten het academische gebeuren. Of het nu de *bioleute*, de vosjes, de chiro/biro, de derpel-crew, de fietskeukenploeg, zomerstraters of anderen zijn, het is heerlijk zo veel waardevolle mensen te mogen kennen.

Ik zou ook graag van de gelegenheid willen gebruik maken om mijn ouders, broer en zus (alsook hun fantastische kroost) te bedanken voor alle steun, het warme nest en de onvoorwaardelijkheid.

Tot slot, Katrijn. Wat begon als een vlotte samenwerking is me intussen het meest dierbaar geworden. Je kent mijn handleiding soms beter dan ikzelf en loodste me naar de eindstreep van dit doctoraat.

Contents

List of Abbreviations

ABC	Approximate Bayesian Computing
ASM	Activated Sludge Model
BATEA	Bayesian Total Error Analysis
BMA	Bayesian Model Averaging
BOD	Biological Oxygen Demand
CDF	Cumulative Density Function
DBM	Data-Based Mechanistic approach
DREAM	DiffeRential Evolution Adaptive Metropolis
DYNIA	DYNamic Identifiability Analysis
EE	Elementary Effect
FDC	Flow Duration Curve
FUSE	Framework for Understanding Structural Errors
GIS	Geographic Information System
GLUE	Generalized Likelihood Uncertainty Estimation
GUI	Graphical User Interface
HBV	Hydrologiska Byråns Vattenbalansavdelning model
HRU	Hydrological Response Unit
IWA	International Water Association
KGE	Kling Gupta efficiency
LH	Latin Hypercube

MAE	Mean Absolute Error
MC	Monte Carlo
MCF	Monte Carlo Filtering
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MSDE	Mean Squared Derivative Error
MSRE	Mean Square Relative Error
NAM	Danish Nedbør Afstrømnings Model
NSE	Nash-Sutcliffe Efficiency
OAT	One factor At a Time
ODE	Ordinary Differential Equation
OED	Optimal Experimental Design
OLS	Ordinary Least Squares
PB	Percent Bias
PDE	Partial Differential Equation
PDF	Probability Density Function
PDM	Probability Distributed Model
REW	Representative Elementary Watershed
RMSE	Root Mean Square Error
RSA	Regional Sensitivity Analysis
RWQM	River Water Quality Model
SA	Sensitivity Analysis
SCE-UA	Shuffled Complex Evolution
SOA	Service-Oriented Architecture
SRC	standardized regression coefficient
SRRC	standardized rank regression coefficient
SSE	Sum of Squared Errors
SWAT	Soil and Water Assessment Tool
TRMSE	Transformed Root Mean Square Error
VHM	Veralgemeend Hydrologisch Model
VIF	Variance Inflation Factor
WSSE	Weighted Sum of Squared Errors
WWTP	Waste Water Treatment Plant

Summary

Mathematical modelling is an important activity in environmental science. Models are used for understanding, prediction, design and optimization. A mathematical model is always a simplified representation of the natural system it attempts to describe. It represents the conceptual thinking about the system processes in a mathematical formulation and translates this into programming code. Once a suitable model has been identified, it becomes a powerful tool for both scientists and engineers.

There is no such thing as the *super-model*, applicable to all situations. Nature is a highly heterogeneous system, which requires a tailor-made description to be effective. The appropriateness of a model structure needs to be sufficiently evaluated taking into account the modelling purpose and the available observational data.

Determining a priori which model structure is most appropriate for a given model application, is a challenging problem. This makes the identification of a suitable model structure an iterative process. Each model structure represents a hypothesis which can be confirmed or rejected by the available observations.

In contrast to this need for adaptation and flexibility, a culture of monolithic model software applications with limited flexibility, is in place. The same legacy models are used over and over again, which led to a vast ignorance among modellers with regard to the appropriateness of the model structure as correct system representation. This resulted in a practice of model parameters fitting instead of model structure identification.

Therefore, the aim of this dissertation is to propose and apply a framework for improved model structure evaluation and identification. The proposed diagnostic approach combines the flexibility to continuously adapt model structures with the means to properly evaluate these alternative representations.

A wide range of existing software environments and frameworks already support flexibility in the model development, but do not always support a rejection framework. To support future research, a minimal set of requirements that needs to be fulfilled is extracted from an analysis of existing tools: (1) the support to al-

ternative representations of the considered processes, (2) the ability to construct alternative configurations, (3) a clear separation between the mathematical and computational model and (4) accessible and modular code implementations.

The model evaluation generalizes the idea of model calibration towards a combined and iterative process of parameter and process (model structural) adaptation. Practical identifiability, both in terms of parameters and model components, is the guiding principle during the evaluation. This means that model structures should contain influential parameters that are not cancelling each other out. In other words, process descriptions should have a clear function that can be consistently identified by the available observations.

The research objective of the modelling exercise needs to be clearly reflected in the performance metrics on which the model structure is evaluated. The central role these metrics have in any kind of model exercise is regularly ignored. The so-called metric oriented approach accommodate the variety of modelling purposes and provide a common denominator for many existing frameworks in literature.

The identification of parameters in complex models is supported by sensitivity analysis. Different methods for sensitivity analysis are audited and implemented as a modular and reusable set of functionalities to support the model evaluation process. This provides a range of tools available to future modellers and initiates tools that can be further developed by and for the environmental modelling community.

In a first application, the identifiability and model calibration of a respirometric model with an additional time-lag component is analysed using the generically implemented tools. The analysis reveals that experimental data for which the ratio between the added substrate and the biomass is high enough needs to be available to properly identify the time-lag component. The appropriateness of the model structure is confirmed and is in line with earlier studies, however subject to the assumptions taken.

In the remainder of the dissertation, lumped hydrological models are studied, describing the relationship between rainfall and runoff.

A first hydrological application studies an ensemble of hydrological model structure alternatives, representing different configurations of the already existing Veralgemeend Hydrologisch Model (VHM). Based on the observed runoff time series, the differentiation of the model structures is not feasible using the chosen set of performance metrics. A lack of parameter identifiability of the individual structures hampers the attribution of model performance to individual model decisions. Hence, there is no added value of creating an ensemble of highly alike structures

when the identifiability of the model structures is not guaranteed. The identifiability of the individual model structures is a necessary condition to compare model structural alternatives and evaluate the correctness of their system representation (hypothesis) in terms of performance.

To enable the interpretation of the appropriateness of model structural decisions when facing unidentifiability, a novel qualitative method for model component sensitivity analysis is introduced. The method enables to make qualitative statements about the relative influence of model structure components towards a chosen performance metric. The application on the ensemble of model alternatives for the case study of the Grote Nete indicated the need for more complexity in the model structure when focusing on low flow conditions.

The last application seeks to diagnose structural errors in two existing lumped hydrological models that are currently applied in operational water management (PDM and NAM). To comply to the requirements of the diagnostic approach, a conversion of both model structures is executed towards a system dynamics representation. It enables the decoupling of the mathematical and computational model and converts both models into a flexible entity supporting alternative model structure configurations.

Besides the implementation in a flexible modelling environment, a standardised matrix representation of lumped hydrological model structures is proposed. The latter provides a common format to communicate about the applied model structure, supporting a reproducible scientific application of lumped hydrological models. The inspiration came from a related scientific field where this is commonly applied and has proven extremely useful. This emphasises the multidisciplinary nature of this work.

To identify the model deficiencies, the DYNamic Identifiability Approach (DYNIA) is applied, a time-variant based method that screens the parameter identifiability as a function of time. In general, similar model performances are observed. However, the model structures tend to behave differently in the course of time. Based on the analyses performed, the probability based soil storage representation of the PDM model outperformed the NAM structure.

In a concluding perspective, some suggestions for an improved development of models and tools for model evaluation are given, based on the gained experiences. In a personal visionary roadmap, the role that open science can have as the engine for collaborative development in the environmental modelling community, is stated.

Samenvatting

Wiskundige modellering is een belangrijk onderdeel van de milieuwetenschappen. Dergelijke modellen worden zowel gebruikt om inzicht te krijgen in een systeem, om voorspellingen te maken en als ontwerp- en optimalisatietool. Een wiskundig model is steeds een vereenvoudigde weergave van het natuurlijke systeem dat het beschrijft. Het model is een conceptuele voorstelling van de systeemprocessen in wiskundige vergelijkingen, die bovendien omgezet worden in programmeercode. Eens een geschikt model opgesteld is, dan wordt het een krachtig hulpmiddel, zowel voor wetenschappers als ingenieurs.

Er bestaat echter geen *supermodel* dat toepasbaar is in alle situaties. De natuur is immers een uiterst heterogeen systeem, waardoor elke onderzoeksvraag nood heeft aan een op maat gemaakte beschrijving. De geschiktheid van de gekozen modelstructuur moet voldoende geëvalueerd worden ten opzichte van het modelleerdoel en de beschikbare geobserveerde data.

Het is een grote uitdaging om het meest geschikte model te vinden voor een gegeven modelleringstoepassing. De identificatie van de geschikte modelstructuur is dan ook een iteratief (aanpassings)proces. Elke mogelijke modelstructuur stelt slechts één mogelijke hypothese voor en die kan door de beschikbare data bevestigd of weerlegd worden.

Ondanks deze duidelijke nood aan flexibiliteit tijdens het opstellen van een model, ontstond er een cultuur van monolithische softwaretoepassingen die slechts een zeer beperkte flexibiliteit toelaten. Hierdoor worden steeds dezelfde welbekende modellen gebruikt, zonder de geschiktheid van de modelstructuur als correcte systeemvoorstelling te evalueren. Dit resulteert in een modelpraktijk van louter het aanpassen van parameters in plaats van eerst de meest geschikte modelstructuur te bepalen.

Het doel van dit proefschrift is dan ook om een raamwerk op te stellen voor een verbeterde model evaluatie en identificatie. De voorgestelde diagnostische aanpak combineert de flexibiliteit die toelaat om de modelstructuur continu aan te passen en de technieken om de alternatieve modellen op een correcte manier te kunnen evalueren.

Hoewel flexibiliteit in de modelontwikkeling ondersteund wordt door een brede waaier aan bestaande softwareomgevingen en raamwerken, wordt het zelden gekoppeld aan de idee dat modelstructuren verworpen moeten kunnen worden op basis van de beschikbare data. In de diagnostische aanpak worden de minimale vereisten voor flexibele modelomgevingen beschreven: (1) het aanmoedigen van het gebruik van alternatieve representaties, i.e. modelstructuren, van het systeem (2) de mogelijkheid om nieuwe alternatieve representaties eenvoudig en transparant op te stellen, (3) de aanwezigheid van een duidelijke scheiding tussen het wiskundige en computationele model en (4) het gebruik van toegankelijke en modulaire implementaties.

De voorgestelde model evaluatie veralgemeent de idee van modelkalibratie (i.e. het aanpassen van parameterwaarden) tot een gecombineerd en iteratief proces van parameter én modelstructuur adaptatie. Praktische identificeerbaarheid, zowel voor parameters als modelcomponenten, is de leidraad tijdens de evaluatie. Dit betekent dat de modelstructuren parameters bevatten met een identificeerbaar effect. Het effect van de parameters op de modeloutput mag elkaar immers niet opheffen. Bovendien moet elke modelcomponent een duidelijk doel hebben dat eenduidig vast te stellen is op basis van de beschikbare data en overeenkomstig de conceptuele voorstelling.

De identificeerbaarheid van parameters in complexe modellen, wordt bepaald met behulp van gevoeligheidsanalyse. In dit proefschrift worden verschillende methodes voor gevoeligheidsanalyse niet alleen uitvoerig en consistent beschreven, maar ook geïmplementeerd als een modulaire en herbruikbare set aan functionaliteiten om het modelevaluatieproces te ondersteunen. Om bruikbaar te zijn voor de verscheidenheid aan modelleerdoelen, werd de implementatie zodanig ontworpen dat de gebruiker op een eenvoudige en snelle wijze de verkozen evaluatiecriteria of performantiecriteria kan opstellen. De implementatie stelt een waaier aan functionaliteiten beschikbaar voor toekomstige modelleerders en kan verder ontwikkeld worden voor én door de modelleergemeenschap van milieutoepassingen.

In een eerste toepassing van de aanpak, wordt de identificeerbaarheid en modelkalibratie van een respirometermodel geanalyseerd met de beschikbare functionaliteiten. Het model bevat een vertragingscomponent om de vertraagde activiteit van de biomassa te beschrijven. De analyse toont aan dat voor experimenten waarbij de verhouding van het toegevoegde substraat tot de hoeveelheid biomassa groot genoeg is, de identificatie van deze vertragingsfactor mogelijk maken. De geschiktheid van de modelstructuur kan onder de genomen assumpties dus bevestigd worden.

Verdere toepassingen in dit proefstuk zijn gericht op hydrologische modellen die de ruimtelijke component niet in rekening brengen (geaggregeerd of *lumped*). Deze modellen beschrijven het verband tussen enerzijds neerslag en anderzijds afstroming (runoff).

Een eerste hydrologische toepassing bestudeert een reeks hydrologische modelstructuuralternatieven, gebaseerd op de verschillende configuraties van het reeds bestaand Veralgemeend Hydrologisch Model (VHM). Op basis van de geobserveerde runoff tijdsreeksen, is het onmogelijk een onderscheid te maken tussen de alternatieve modelstructuren op basis van de gekozen set van evaluatiecriteria. De oorzaak van deze tekortkoming is waarschijnlijk het gebrek aan parameter identificeerbaarheid van de individuele modelstructuren, waardoor de performantie van een modelstructuur niet eenduidig toegekend kan worden aan de individuele modelcomponenten. Het heeft dus geen zin om een reeks aan zeer gelijkaardige modelstructuren op te stellen als de identificeerbaarheid ervan niet gegarandeerd wordt. Deze identificeerbaarheid blijkt een belangrijke voorwaarde om modelstructuuralternatieven op basis van hun performantie te kunnen onderscheiden.

Om ondanks identificeerbaarheidsproblemen, toch een uitspraak te kunnen doen over de geschiktheid van modelcomponenten, wordt een kwalitatieve methode voor sensitiviteitsanalyse voorgesteld, gericht op modelcomponenten. De vooropgestelde methode maakt het mogelijk om de relatieve invloed van de verschillende modelcomponenten op de verkozen evaluatiecrieteria weer te geven. De toepassing van deze methode op de reeks aan modelstructuren van het VHM leidt tot concrete voorstellen voor modelstructuuradaptatie, zoals de noodzaak tot een complexere modelstructuur om condities waarin weinig water door de riviers stroomt goed te kunnen modelleren.

In de laatste toepassing in dit proefwerk worden twee bestaande *lumped* hydrologische modellen (PDM en NAM), die in het huidige operationele waterbeheer gebruikt worden, onderzocht met als doel structurele fouten op te sporen. Om overeenkomstig de diagnostische aanpak te handelen, worden beide modelstructuren omgezet in een systeemdynamische voorstelling. Deze aanpassing ontkoppelt het wiskundige en computationele model en zorgt voor een flexibele implementatie die het gebruik van alternatieve modelstructuren ondersteunt.

Vervolgens wordt een gestandaardiseerde matrixvoorstelling voor *lumped* hydrologische modellen voorgelegd en toegepast op ondermeer PDM en NAM. Deze matrixvoorstelling zorgt voor een eenduidige voorstelling van in de literatuur beschikbare modellen. Hierdoor wordt de communicatie rond modelstructuren vereenvoudigd en wordt een belangrijke stap gezet in de richting van reproduceer-

baarheid omtrent modelstudies uitgevoerd op basis van *lumped* hydrologische modellen.

Om vervolgens de modeltekortkomingen van NAM en PDM op te sporen, wordt de DYNamic Identifiability Approach (DYNIA) toegepast. In deze methode wordt de parameter identificeerbaarheid nagegaan op de verschillende tijdstippen van de simulatie. In het algemeen, vertonen beide modellen een vergelijkbare performantie. Toch blijkt uit de uitgevoerde analyses dat de modelstructuren zich anders gedragen op verschillende tijdstippen. Er kan gesteld worden dat de op probabiliteiten gebaseerde bodemopslagvoorstelling van het PDM model, het NAM model overklast voor de specifieke toepassing.

In de afsluitende perspectieven worden vanuit de eigen ervaring rond de ontwikkeling van modellen en tools voor modelevaluatie, enkele suggesties gedaan ter verbetering. Een visie wordt geschetst van hoe open wetenschap de drijvende kracht kan vormen voor een gezamenlijke ontwikkeling in de modelleringswereld.

PART I

INTRODUCTION

CHAPTER 1

Problem statement, objectives and outline

In a fast developing world with an ever rising population, the pressure on our natural environment is continuously increasing. The growing world population associated with an expanding industrial activity, intensified agriculture and increased competition for land and resources is causing multiple environmental issues.

The large variety in environmental issues resulted in a wide range of scientific disciplines focussing on different components of the natural environment and environmental technologies. Notwithstanding the huge differences amongst the environmentally oriented research disciplines and their respective focus, modelling has become an important activity in environmental science in general. Models are used for understanding, prediction, design and optimization.

As one of the above mentioned environmental issues, our natural water resources are under stress, leading to a poor water quality of streams, rivers, lakes and seas. Besides, both water scarcity and floods threaten humans all over the planet. The specific reasons and mechanisms causing these threats differ amongst different spatial and temporal scales and as such, insight in the driving mechanisms is essential in order to mitigate these problems. During the last decades, a model-based approach has become an essential part of scientific research in continuous interaction with the increased capabilities of measurement devices.

A (mathematical) model is understood as a simplified representation of the natural system it attempts to describe (Refsgaard, 2004; Gupta et al., 2008). As such, it represents the conceptual thinking about the system functioning in a mathematical formulation and translates this into programming code. In other words,

the implemented model can be regarded as a set of hypotheses of the underlying mechanisms, which can be either confirmed (or at least considered reliable) or rather falsified based on the ability to correspond to real-world observations (i.e. data).

The real-world is a highly diverse system that is studied at a huge range of both temporal and spatial scales. Different research questions require an alternative focus on a specific segment of the environmental system, leading to different mechanisms to conceptualize and describe. In this respect, the highly heterogeneous environmental systems request for a tailor-made approach to be effective. In other words, there is no such thing as the *super-model* or *one-fits-all* model. On the contrary, considering the conceptual properties of a model, a set of potentially suitable models for each problem at hand do exist and some of them will be fit for purpose.

This leads to two main **challenges**. First, the capacity of building and implementing different models that possibly are suitable and fit for purpose. Secondly, the ability to test, compare and diagnose these implemented models in order to evaluate the properties and performance of the individual model structures and to come up with an appropriate model (and eventually, an ensemble of models) supported by the available real-world observations. As one can expect, this will be an iterative process, since failure of each of the proposed models will lead to new proposals based on the learned shortcomings. To make this useful to practitioners, this learning process needs to be transparent, fast and usable.

Progression has been made with respect to these two challenges. A *plethora* of models and modelling frameworks to construct models exists in all sub-fields of environmental science. Moreover, a wide range of methodologies has been developed to evaluate model performance. With an ever increasing computational capacity, this leads to enormous opportunities. However, at the same time a huge conservatism and *default-settings* practice does exist in the application of model-based analysis in contradiction to the required tailor-made approach. When it comes to practical applications, the same legacy models are used over and over again (1) ignoring their uselessness/usefulness, (2) using the same (rather minimalistic) model evaluation criteria and (3) only reporting positively about the obtained modelling results. This painfully demonstrates the gap between modelling community *common practice* and potential *best practices*. The lack of accessibility and portability, the closed source nature of many modelling software platforms, the lack of programming skills of environmental engineers, ignorance or sheer protectionism are just some of the reasons preserving that gap, despite many well-intentioned initiatives.

This dissertation is not claiming to overcome this gap, but rather explores the possibilities on how to improve current model-based analysis. As such, it does not provide a classical research hypothesis driven insight or an application driven narrative, but rather a methodological exploration illustrated on specific applications. By accepting the method of multiple working hypotheses (Chamberlin, 1965; Kavetski and Fenicia, 2011) and by applying model-based analysis as a learning by failure approach (Beven et al., 2007), it is aimed to provide response to current model parameter fitting practices commonly encountered. This approach requires a flexible implementation of model structures and diagnostic techniques for model structure evaluation. The dissertation aims to develop methodologies which should pave the way to an improved model diagnostic approach and more reliable models, within a more transparent and reproducible scientific practice.

1.1 Problem statement

An imbalance exists between the scientific research on the identification of an appropriate model structure, compared to the applications of model analysis methodologies on an already predefined model structure. Figure 1.1 illustrates this imbalance by showing the relative amount of papers in Web of Science resulting from a search on the defined term as topic. Results were restricted to the research areas 'Water resources' and 'Environmental sciences/Ecology'. Only around 11% of the papers are handling one of the topics 'model identification', 'model discrimination', 'model selection' or 'structure characterisation', whereas 44% are about 'sensitivity analysis', 13% about 'uncertainty analysis' and 32% about the topic of model calibration. The latter is an aggregation of the counts on the search term 'model calibration' itself and results provided by the search terms 'parameter optimization', 'parameter estimation' or 'inverse modelling'.

Notwithstanding the diversity of currently existing models and modelling frameworks, **the identification of the most appropriate model structure for a given problem remains an outstanding research challenge**. Model structure evaluation based on aggregated performance measures do provide a general assessment about the goodness of fit, but do not provide information about why a particular model structure performs better or worse. Tools and sound procedures to diagnose a model structure in order to identify deficiencies are clearly underrepresented in literature. Note that the state of development can be very different in various fields. Exchange of methods and procedures between disciplines is also relatively limited.

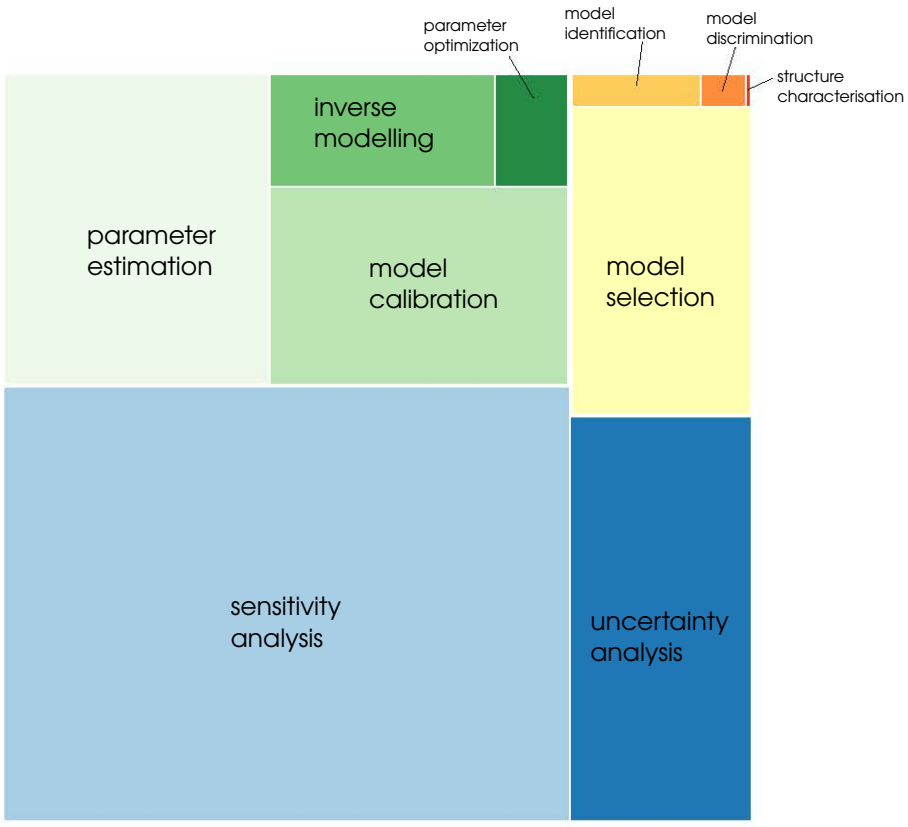


Figure 1.1: Treemap visualisation of the relative amount of papers (represented by the area in the graph) enlisted by Web of Science when querying for the specified search term as topic for the research paper within the research areas 'Water resources'and 'Environmental sciences/Ecology'on the entire historical database.

1.2 Research objectives

In view of the problem stated above, the general aim of the dissertation is to **improve current practice of model structure comparison and evaluation by making individual model decisions explicitly testable**. To achieve this, the following sub-objectives were defined, thematically divided into 4 main themes (the objectives are numbered and tagged according to the specific theme):

1. Definition of a **diagnostic approach (D)**, supporting an improved model structure evaluation

- **Objective D.1**: *Obtain insight in the current lack of coherence within the field of environmental modelling, leading to conservative practices.*

  The rather limited research towards model structure identification in contrast to the numerous work reporting the application and calibration of existing model structures, is apparent. Instead of questioning the model structure itself, the model is recycled to address new problems by tuning the parameters only. The aim is to understand the driving factors triggering this evolution.

- **Objective D.2**: *Define the requirements of an improved diagnostic approach for model-based analysis.*

  Based on the insight provided by Objective D.1, the aim is to propose an alternative general diagnostic framework for model structure evaluation.

2. Improve current practice in terms of **model evaluation tools (E)**

   - **Objective E.1**: *Propose a metric oriented approach as a common denominator for the current plethora of existing model evaluation tools.*

     A wide set of methodologies for model analysis does already exist and numerous procedures are described in literature. Notwithstanding the diversity of existing methods, the aim is to find the common building blocks of these methods and illustrate how the chosen metric is the central element in most of these methods.

   - **Objective E.2**: *Facilitate the application of sensitivity analysis for model evaluation by providing an open and extensible implementation*

     Implementations for performing sensitivity analysis are scattered, not provided together with publications, poorly documented and diverse in the algorithmic choices. The aim is to provide an implementation of some existing methods for sensitivity analysis that is open, extensible and accommodated with documentation of the code.

3. Improve current practice in terms of **model structure development (S)**

   - **Objective S.1**: *Define a general set of requirements for model structure development that supports model structure evaluation*

     Providing alternative model structure configurations is supported by modelling environments that provide flexible model development. Still, this does not automatically mean they support the diagnostic approach proposed in this dissertation. The aim is to define the minimal require-

ments for (flexible) modelling environments to support the diagnostic approach.

- **Objective S.2**: *Development of an implementation independent and standardised model structure description for hydrological models, making communication about hydrological model structures explicit and transparent.*

  Lumped hydrological rainfall runoff models are a group of environmental models that are well-known, for both prediction (e.g. flood events) as well as integrated modelling applications. In essence, this group of models can be conceived as a set of ordinary differential equations, representing the mass balances of interlinked reservoirs. Whereas this supports maximal flexibility, the reporting in literature is mostly specifying one specific model configuration, represented by an acronym. The latter does not support a clear and transparent communication of the applied model structure, making comparison cumbersome. The aim is to overcome this issue by providing a summarized matrix representation of a lumped hydrological model.

4. **Apply and extend (A)** the current set of model evaluation tools to support the evaluation of individual model decisions

   - **Objective A.1**: *Illustrate the metric oriented approach by performing an identifiability analysis on a respirometric model.*

     To illustrate the idea of a metric oriented approach, a respirometric model is used to check the identifiability of the parameters and perform a calibration to real-world observations.

   - **Objective A.2**: *Assess the usefulness of parameter optimization to differentiate model structure decisions within a flexible model environment.*

     When seeking an optimal model structure amongst an ensemble of models, the most straightforward option is to define a set of performance metrics and compare the metrics among the members of the ensemble, choosing the best performing one. The question now arises, if this approach could be used to differentiate between the members of a flexible model environment, where these members do have common components. The aim is to check the difference in performance for an ensemble of model structures derived from the Veralgemeend Hydrologisch Model (VHM), a lumped hydrological model.

- **Objective A.3**: *Extend current sensitivity analysis to reveal the effect of changes in the model structure and evaluate specific model structure decisions*

  A performance metric on itself does not directly provide information about *why* a certain model structure is better or worse. In order to gain insight in the reasons why a model structure is performing well and link it to individual model processes (components), alternative information is sought. Sensitivity analysis is a well known technique to link the influence of model parameters with the predicted output, but it does not provide information about the influence of individual model components. The aim is to extend the usage of sensitivity analysis to the level of model components in order to derive information about the usefulness of model components for a specified model objective.

- **Objective A.4**: *Use a time-variant based evaluation of model structures to identify model structure deficiencies.*

  Within a model structure definition, model parameters are supposed to have constant values (within some uncertain ranges). Parameter values that should be changed in function of time to properly represent the observations, indicate a missing aspect in the model formulation. Starting from this idea, the aim is to identify model deficiencies by actively allowing the parameters to vary in function of time.

1.3 A road-map through this dissertation

The dissertation consists of one introductory part (Part I), three main parts (Parts II to IV) and a concluding Epilogue. The different parts are composed of several chapters. This structure, as well as the interdependencies of the Parts and chapters is visualized in Figure 1.2 and briefly discussed here.

Part I: The diagnostic framework

Chapter 2 provides a more elaborate insight into the central problem statement of the dissertation, providing an answer to **objective D.1** by identifying and describing some main drivers leading to the conservative practice of model fitting as parameter tuning instead of model structure identification.

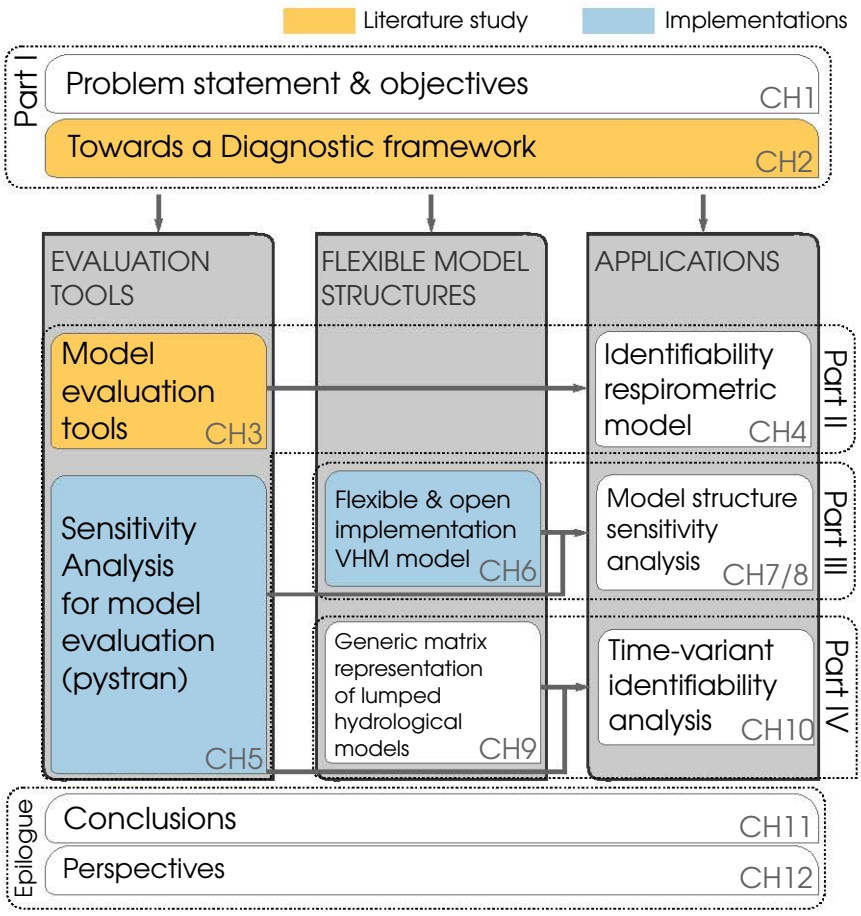


Figure 1.2: Roadmap of the dissertation, providing the interrelations in between the different chapters.

Starting from these observations, an alternative general framework is proposed in chapter 2 that provides the conditions to overcome this conservatism as defined by **objective D.2**. In the last section of chapter 2, the necessary requirements for flexible model environments to accommodate the diagnostic approach are discussed as part of the diagnostic approach, answering **objective S.1**.

Part I provides a methodological background based on literature study and should be regarded as the general setting in which the remainder of the dissertation is embedded. Any type of reader is encouraged to read this part. For experienced modellers, it provides a critical reflection on the current practice and how this could be altered, whereas it can help newbie modellers in better understanding the iterative cycle of a model based analysis.

Part II: Model diagnostic tools

Part II consists of chapter 3, 4 and 5 and focuses on tools for model analysis and evaluation. Chapter 3 starts from the observation that many methodologies exist in parallel, but actually rely on a set of common building blocks due to the characteristics of environmental models. Instead of focusing on specific algorithms, the chapter deals with **objective E.1** by putting the choice and construction of the metric central. The chapter can be regarded as a general literature overview of existing tools for model evaluation from the point of view of metric construction, without going into detail on specific algorithms. To illustrate the metric oriented approach (**objective A.1**), a first case study on a respirometric model is performed in chapter 4. The chapter illustrates how complementary information can be extracted by using different aggregated (performance) metrics of the model output.

Next, a detailed description on a subset of methods is provided in chapter 5, with particular focus on sensitivity analysis. The latter enables to verify the influence of input factors (e.g. parameters) with respect to the modelled output, which is of particular interest to assess model structure behaviour. To overcome the lack of code documentation and transparency of existing implementations, the implemented methods were collected in a Python package, called pystran. This facilitates the future application and extension of methods for sensitivity analysis, as defined by **objective E.2**.

Chapter 5 provides the theoretical background and describes in detail the functioning of the implementations of pystran. Some of the methods are used in the subsequent parts of the dissertation, but readers familiar with these well-known methods could safely skip this chapter. Readers who are using the Python implementation will value this chapter to get more insight in the theoretical background of the implementations. For the source code documentation, the reader is referred to the online documentation[1].

Part III: Comparison of hydrological model structure alternatives

The abilities for model identification within a flexible modelling environment are investigated for a particular hydrological model, called VHM. In chapter 6, a limited set of alternative representations of VHM is presented by adapting the original model structure. The resulting model structures are considered as alternative

[1] http://stijnvanhoey.github.io/pystran/

system representations of the study catchment and this set of model structures provide the experimental conditions for chapter 7 and chapter 8.

Chapter 7 compares these model structures in terms of their performance, hereby addressing **objective A.2**. Furthermore, to extract information about model structure decisions within the ensemble independent from an optimized parameter set, chapter 8 aims to extend the classical usage of sensitivity analysis. Instead of evaluating the effect of parameters on the model output, the effect of individual model structural decisions is assessed to meet **objective A.3**.

Overall, the lack of identifiability of the individual model structures, the related impossibility to distinguish the model structures and the difficulty to properly distinguish the mathematical and computational model for the set of model alternatives provided by VHM, lead to the objectives dealt with in the last part of the dissertation.

Part IV: Diagnosing structural errors in lumped hydrological models

Chapter 9, addresses **objective S.2**, striving to provide an implementation independent way to communicate about model structures that fulfil the requirements as defined by **objective S.1**. The matrix representation proposed is inspired by its use in other scientific fields and adopted to enable the description of a wide range of lumped hydrological models.

The identification of model deficiencies is an essential step to propose model adaptations. Chapter 10 focuses on two specific lumped hydrological models that are commonly used in operational water management in Flanders. It seeks to meet **objective A.4**, a time-variant model structure evaluation. The evaluation is based on a time variant investigation of the model structures and screens the parameter sensitivity and identifiability as a function of time.

Epilogue: Conclusions and perspectives

In a closing Epilogue, the main conclusions of this dissertation are provided in chapter 11. Some personal reflections and perspectives are summarised in chapter 12, framing the necessary further steps in terms of model development and diagnostic tools in the context of reproducible and open research.

CHAPTER 2

Towards a diagnostic approach in environmental modelling

2.1 Introduction

When explaining processes and phenomena of nature, scientists make observations or collect experimental data, after which patterns and regularities are sought, mostly supported by statistical analysis. However, statistical correlations on their own do not constitute understanding, neither causality (this does not mean that a correlative diagnostic cannot provide system understanding (Gupta et al., 2008)).

When underlying principles can be identified from which an explanation of the observed patterns and regularities can be derived, this leads to the formulation of a scientific theory capable of making predictions (Shou et al., 2015).

A (mathematical) model structure is in essence one way of formulating such a scientific theory, which can be adapted, extended or falsified by new observations. In essence, all environmental models represent simplified representations of the real world, so proper evaluation and testing is essential (Kavetski and Fenicia, 2011).

Environmental modelling embrace a wide range of scientific fields, which is also reflected in the range of existing model types, going from pure data-based models (essentially linear regression is a data-based model) to detailed models describing complex systems with all its interactions.

In this chapter, first, the type of model representation applied in this dissertation is introduced along with some essential concepts of modelling literature to set the stage. Subsequently, current pitfalls of environmental modelling are identified and discussed. They provide the motivation to the proposal of a diagnostic framework, which is explained in the last section of this chapter.

The aim of this chapter is also to provide some clarification in the wide diversity of nomenclature and terminology used in environmental modelling. It does not have the ambition to provide a full overview, but assists in understanding and contextualizing some central issues to support future modellers.

2.2 Mathematical model representation

Any model representation starts with the delineation of a system together with the system boundaries for which the model applies. The term system can be interpreted widely (Meadows, 2009). It is any entity in which variables of different kinds interact and produce observable signals. The defined domain of the system is a direct function of the research question. It can represent a lab-controlled system (e.g. bio-reactor), a specific element of the environment (e.g. soil compartment, river stretch), an environmental entity (e.g. catchment, habitat)...

In environmental science, continuous (in both space and time) aspects of systems are usually studied, and for complex systems traditional, equation-based approaches are typically most convenient (Claeys, 2008). Hence, focus of this dissertation is specifically on continuous dynamical systems described by deterministic models in the form of a set of (possibly mixed) differential and algebraic equations, using the following notations (Donckels, 2009):

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \boldsymbol{f}(\mathbf{x}(t), \mathbf{y}_{\mathrm{t,in}}(t), \boldsymbol{\theta}, t); \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2.1}$$

$$\hat{\mathbf{y}} = \boldsymbol{g}(\mathbf{x}(t), \mathbf{y}_{\mathrm{t,in}}(t), \boldsymbol{\theta}, t) \tag{2.2}$$

with $\mathbf{x}$ representing a vector of time-dependent (internal) state variables, $\boldsymbol{\theta}$ the vector of k model parameters, $\mathbf{y}_{\mathrm{t,in}}$ a set of forcing variables (in system dynamics regularly expressed as forcing $\mathbf{u}$) and $\hat{\mathbf{y}}$ represents a vector of observed response variables that are function of the state variables $\mathbf{x}$. The algebraic part of the model can be interpreted as a set of derived variables as well as any kind of aggregation function applied on the model state variables, both referred to as the variables of

interest. Hence, $\boldsymbol{g}$ can also simply act as a selector, selecting those (internal) state variables that are actually observed (section 3.3).

All the variables are functions of time t. The system boundaries for which the model is developed are chosen in function of the research objective while taking into account that the fluxes through the boundaries of the defined system, i.e. forcing variables $\mathbf{y}_{t,\text{in}}$, can be easily quantified (as far as possible).

A model simulation is the act of solving the model for a given set of model parameters, initial values $\mathbf{x}_0$ and specified forcing variables. In many cases, solving the mathematical model is not feasible analytically and the application of numerical techniques (solvers) is required (Donckels, 2009).

Since environmental systems are poorly defined, the investigator is ignorant of the 'real' structure and the (non-linear) relationships in between the system variables are unknown (model structural uncertainty). Moreover, available observations are always corrupted to some perspective (data uncertainty) and in many cases insufficient to identify the required model structure (i.e. set of equations) unambiguously. Considering these uncertainties, the task of developing a proper model structure is challenging and of vital importance. In essence, all environmental models represent simplified hypotheses of the real world functioning and these hypotheses require rigorous construction, implementation, evaluation and testing.

2.3 Model structure identification

The task of defining a proper model structure for the problem at hand has been referred to as a challenging problem in the previous section. Different sources in literature provide guidelines and suggestions about the process of model building (Dochain and Vanrolleghem, 2001; Sivapalan et al., 2003; Refsgaard, 2004; Gupta et al., 2008; Fenicia, 2008; Gupta et al., 2012).

There is no agreement on an existing general framework for model building and there is no consensus on the steps to undertake. However, three important stages (levels) can be identified. A first stage is the translation of the real system to an abstract representation of how the system is interpreted, referred to as the conceptual model:

Definition 2.1. *A **conceptual model** is the abstract representation of a real system as a set of interacting processes by the ideas on its constituents and functional relationships*

Some authors make a further distinction in between perceptual and conceptual models (Beven, 2000; Fenicia, 2008), but this merely obscures the terminology (Gupta et al., 2012).

Hence, in contrast to purely data-based approaches, the model development is imposed by prior knowledge (or hypotheses) about the system studied. A second important stage is the translation of the conceptual model into a mathematical model, as represented by Equation 2.1 and defined as (after Kavetski and Clark (2011)):

Definition 2.2. *A **mathematical model** defines the set of initial and boundary conditions of the system, the forcing and response variables, the parameters and the equations to represent the processes defined by the conceptual model*

It is important to discriminate the mathematical model from a third stage which is the computational model, defined as follows:

Definition 2.3. *A **computational model** is the computational implementation of the mathematical model, specifying the numerical or analytical formulation used to solve the governing model equations.*

The three stages above do not directly link to a consecutive set of actions or unilateral workflow, but are part of a **continuous iterative process** of adaptation (Carstensen et al., 1997; Dochain and Vanrolleghem, 2001). Any kind of **model evaluation** can drive this iterative process towards an adapted (improved) representation (i.e. hypotheses). Hence, model evaluation can be regarded as a comprehensive term for modelling techniques that provide insight about the model and its performance.

The included model parameters are generally not directly known and need to be adapted to improve the alignment of the model and the observations. Hence, model calibration is an essential part of the evaluation process and is defined as follows:

Definition 2.4. ***Model calibration** is the adjustment of parameter values that lead to an improved agreement of model results with observed data in which the agreement is expressed in any kind of qualitative and/or quantative metric.*

The ultimate aim of the iterative learning cycle is to identify a model structure that can be successfully applied, so the overall process is also referred to as *model identification* (Dochain and Vanrolleghem, 2001):

Definition 2.5. *The goal of* ***model identification*** *is to find and calibrate a model for the system under investigation that is adequate for the intended purpose*

This definition takes also into account the selection in between different model structures. *Model selection* (Dochain and Vanrolleghem (2001) also call this structure characterisation) follows from the situation in which it is very difficult or even impossible to further discriminate among a set of model structures using the available observations (hydrologists like to refer to equifinality (Beven, 2006)). Techniques to design new experiments to facilitate this discrimination are referred to as Optimal Experimental Design (OED) for *model discrimination* (Donckels, 2009; Asprey and Macchietto, 2000). However, as many environmental systems cannot be controlled, this is not always feasible and one has to work with the available observations. Furthermore, *system identification* is the term that has been used in the control community, which can be more regarded as a purely data driven approach where focus is on the fit itself (independent from how it has been achieved).

2.3.1 Top-down versus bottom-up

The model development approach of this dissertation is made by explicitly considering a conceptual representation (hypothesis of the system), which is not used in a data driven (machine learning) approach.

The distinction is not always so clear and provokes lots of discussion (Sivapalan et al., 2003; Kavetski and Fenicia, 2011; Beven, 2002; Fenicia, 2008; Sivakumar, 2004; Refsgaard, 2004). The underlying methodology for model construction is also divided in between a *deductive* approach (also referred as upward, bottom-up or reductionist approach, theoretical, mechanistic, white box) and an *inductive* approach (or top-down, downward, data-driven, empirical, black box).

However, most models in environmental science (ecological, water quality, hydrological...) are a mixture of empirical and physical descriptions describing different subphenomena (i.e. process descriptions). As such, most (if not all) models are *grey-box* models, representing different processes and their interconnections. Empirical relationships (miniature data-driven models) for individual processes have always been used and are (deeply) nested into a wide range of 'physically based' environmental models (Sivapalan et al., 2003). By accepting this as common, the communication would be less scattered and obscure.

The conceptual representation of the system is represented by a set of interacting processes (see Definition 2.1). First of all, there is no reason to make statements

about physically based or empirical on a model structure level, only on the process level. Secondly, throughout time, the process description can be altered from an empirical relation towards a physical representation when more information (data and/or knowledge) is available. The latter consists of a set of process descriptions, making it a hierarchical set of empirical and physical processes. Therefore, the term (hierarchical) **process based** approach is used to define this type of modelling. It distinguishes itself from a pure data-driven approach by the explicit consideration of interacting processes (hypothesis), without making statements about 'physical' or 'empirical' of the individual process descriptions.

2.3.2 Model validation

Both in terms of model calibration and model identification, the issue of sufficiency and adequacy arises. In other words, *when is a model calibrated?*, respectively, *when is the model appropriate?* At a certain point, the proposed model is *validated*, either leading to the conclusion that the provided model is fit for the purpose and accepted as such, or improvement is needed. However, this is not different from the model identification process itself. Model *validation* is however different as it is mostly linked with the evaluation of the model to an **independent (new) set of observations**, not used in the identification process.

The most well-known practice is a split-sample approach, dividing the observation in a set for model identification and a set for validation. When the model performance declines profoundly when moving to the validation set, an indication is given about the malfunctioning of the model structure. The predictive capability of a model must be evaluated against independent data (Refsgaard, 2004). Still, a model rejection definition is needed with respect to the independent set of observations. Actually, the question *when is the model appropriate?* is only passed on. The decision is generally based on expert-judgement, i.e. subjective (giving rise to typical statements like: *in general, the model fits the data well*). Formal definitions are not widespread in literature. Abbaspour (2005) defines a combined parameter-estimation and uncertainty definition for adequate calibration. However, both are very dependent on the used methodology and its related assumptions (Cierkens et al., 2012), making it subjective as well.

Model validation is subject to discussion as well (Dochain and Vanrolleghem, 2001; Oreskes et al., 1994; Konikow and Bredehoeft, 1992; Refsgaard, 2004). The main argument defines that a hypothesis (in this case a model representation as hypothesis) can never be proven to be generally valid, but may in contrary be falsified by just one example (Oreskes et al., 1994). For example, regardless of how often

we see a white swan, we cannot conclude that all swans are white. However, a single observation of a blue swan would lead to the rejection of this hypothesis, i.e. it can be falsified (Wagener et al. (2001b) according to Magee (1973)). This directly refers to statistical testing and the falsification idea of Popper (1959)(section 2.5.1).

Models are indeed only representations, useful for guiding further study but not susceptible to proof. Still, one can evaluate whether it is appropriate for its intended purpose (engineering approach) (Kuczera et al., 2010). This means a model needs to be tested for tasks it is specifically intended for and should only be used with respect to outputs that have been explicitly validated (Refsgaard, 2004). Therefore, transparency in the evaluation process is essential.

Finally, validation is sometimes also called *verification*. However, the latter generally refers to checking the implementation of the model and necessary numerics, i.e. the computational model (Refsgaard, 2004).

2.3.3 Identifiability

Within the process of model identification, a main indicator for model deficiency is the inability to find a unique parameter combination that is able to describe the data most appropriately. A lack of identifiability can be related to the model structure itself (structural identifiability) or to the quantity and quality of the experimental data (practical identifiability) (Vanrolleghem et al., 1995):

- **Structural identifiability** of a model structure is examined under the assumption that perfect or error-free measurements are available for the response variables and is purely based on the mathematical model itself.
- **Practical identifiability** determines whether the available data is sufficiently informative to identify the model parameters. It investigates if the available observations are informative enough to give the parameters unique and accurate values.

A parameter that is practically identifiable is also structurally identifiable but not vice versa. For linear models, the derivation of structural identifiability is well-developed and a variety of methods do exist. However, for non-linear models, the application is less straightforward and requires direct manipulation of the mathematical model by symbolic software, which is not so feasible in many environmental applications (section 2.4.4) (Dochain and Vanrolleghem, 2001).

Hence, the determination of the practical identifiability is essential. The identifiability can be quantified in different ways, but measures generally consist of an evaluation of the **sensitivity of the output to the parameter** and on the **dependency of the parameter** on other parameters, i.e. interaction (De Pauw et al., 2008). A model parameter that is highly influential towards the model output and which effect is not cancelled out by the effect of changes of other parameters can be regarded as identifiable.

At the same time, it clarifies the need for sufficient data because data provides the dynamic conditions for the simulation on which the identifiability analysis is performed. Data availability when parameters are not influential does not provide any added value. In other words, adding complexity to a model structure without the data to identify the parameters does not make sense. From this, the application of OED originates, proposing new experiments to increase the identifiability (Donckels, 2009).

2.4 Conservatism in environmental modelling

Modelling is a multi-disciplinary field, confronting the knowledge of environmental processes with sub-domains of mathematics and computer science. However, environmental scientists are mostly not trained in computational or mathematical science and have typically an environmental domain specific background. Hence, the adaptation of modelling concepts is sometimes very fragmented and ad-hoc. Specific modelling methodologies are favoured within scientific fields, mainly because of being most appropriate, but regularly just pure out of conservatism, tradition and ad hoc training.

Notwithstanding the in general common mathematical blueprint (Equation 2.1), a sprawl of modelling environments, technologies and practices are communicated, giving rise to a lack of coherence in the scientific modelling field. In addition, terminology amongst disciplines is different, causing barriers for interdisciplinary exchange. This hampers researchers in the selection of the methodologies and is obscuring the interpretation of the individual methodologies (Carstensen et al., 1997; Montanari, 2007; Refsgaard, 2004). However, when focusing on the literature, a lot of similarities can be identified amongst them (section 3.2).

Furthermore, notwithstanding the continuous progression that is made in all of the scientific fields, the adaptation towards new technologies is hampered. Practitioners are not able to easily employ new technologies, leading to conservative practices.

In this section, some major issues of conservatism within the environmental field are identified and discussed to gain more understanding. This will enable the proposal of an alternative approach that can support the transition towards more integrated and adaptive modelling practices amongst different scientific fields.

2.4.1 Incoherent terminology

As illustrated in section 2.3 (a sigh of desperation while reading that set of definitions and terminology, is completely acceptable), the usage of the same modelling terminology is not agreed upon between communities and even not within model communities. Different authors highlight the lack of coherency and clarity in modelling terminology (Montanari, 2007; Carstensen et al., 1997; Refsgaard, 2004; Gupta et al., 2012; Sivakumar, 2008). This hampers the communication and leads to misunderstanding which can result in wrong expectations and undermines the confidence of stakeholders. Too strict modelling guidelines can lead to a limitation of the scientific progress, but it is important for practitioners to transfer scientific good practices.

Montanari (2007) also refers to this lack of a systematic approach, limiting the scientific transfer. However, as stated by Refsgaard (2004), the confusion on terminology and the lack of common terminology itself is one of the reasons hampering the establishment of generally acceptable modelling guidelines. Model calibration, an essential step in model development (section 2.3), is a well understandable example. Amongst different communities, the adjustment of parameter values in order to improve the model fit using a specific data set, is known. However, depending on the specific research field, this is referred as *model optimization, parameter calibration, model calibration, inverse modelling, parameterization or parameter estimation.*

An extensive glossary of modelling terminology was provided by Carstensen et al. (1997) within the water quality community and a comprehensive terminology for model credibility was presented by Schlesinger et al. (1979) as a report to the general membership of the *Society for Modeling & Simulation International.* However, these are just two of the many societies active in modelling (cfr. International Water Association (IWA), European Geoscience Union (EGU) amongst others). The coverage and comparison of model adequacy testing among the groundwater, unsaturated zone, terrestrial hydrometeorology, and surface water communities is seldom seen (Gupta et al., 2012). A good glossary of modelling terminology is provided by Carstensen et al. (1997) and later published in Dochain and Vanrolleghem (2001).

Claeys (2008) distinguishes in between sub-domains of environmental science that can be considered as mature, accepting a set of methodologies and standardized models and sub-domains that lack this consistency and *lingua france*, having a long way towards consolidation of ideas and procedures. In this respect water quality management is considered mature, accepting standards (cfr. the Activated Sludge Model (ASM) series and River Water Quality Model (RWQM) (Claeys, 2008)) and providing good modelling practice manuals as an outcome of IWA community *specialist groups*. On the other hand, the hydrological domain does have some widely used models, but no general accepted guidelines or procedures. However, it can be doubted if this is only a growing process towards maturity or rather the inherent characteristics on the so-called *uniqueness of place* (Beven, 2000), (section 2.4.6). Moreover, the usage of benchmarks to assess the added value is also known in hydrological forecasting (Pappenberger et al., 2015) and land models describing biophysical processes (exchanges of water and energy) and biogeochemical cycles of carbon, nitrogen, and trace gases (Luo et al., 2012). Recently, a vocabulary to communicate in a standardised way about hydrological modelling observations has been proposed (Horsburgh et al., 2014).

It is evident that unclear terminology limits the transparency and reproducibility of the scientific process, which is an essential condition for scientific progress. This becomes even more considerable in a multidisciplinary field as environmental modelling, where a wide range of expertises is needed to push knowledge forward. Openness and transparency on every level are essential to really define what can be considered as added value.

2.4.2 Quest for a detailed and complex description

In section 2.3.1 insight is given in the bottom-up model construction approach focusing on process understanding by continuously adding more detail and complexity to model descriptions. The central problem with the increased detail is not the creation and implementation of these descriptions, but the infeasibility of the application when data availability is insufficient. The latter makes it impossible to calibrate the increasing number of parameters (Sivakumar, 2004; Beven, 2002; Sivakumar, 2008). Kirchner (2006) argues as follows:

> I argue that scientific progress will mostly be achieved through the collision of theory and data, rather than through increasingly elaborate and parameter-rich models that may succeed as mathematical marionettes, dancing to match the calibration data even if their underlying premises are unrealistic.

In other words, the models do have sufficient degrees of freedom (parameters) to provide acceptable simulation results, certainly when *acceptable* is not too tightly defined (Beven, 2000).

This gives rise to a paradox for model selection: the more complexity is added, the easier one would expect it is to make a distinction between model structures. However, the increased degrees of freedom makes it more difficult to discriminate models structures, since they each have more options (parameter combinations) to provide acceptable results. The latter makes it harder to define which model provides the *right answers for the right reasons* (Kirchner, 2006). This discussion is of course always conditioned on the number and quality of available observations. More complex models need more data, both in terms of forcing variables and observations to evaluate the performance (Sivapalan et al., 2003). Bottom line is that the available observations delimit the detail that can be represented and if someone aspires a more detailed description, proper observations should be collected (Figure 2.1).

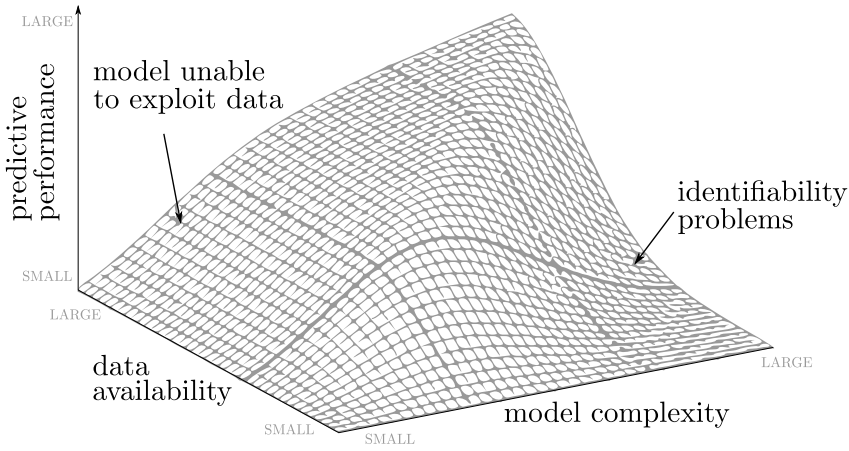


Figure 2.1: Illustration of the balance between the available data and the model complexity. Too much detail of the process description leads to identifiability issues when insufficient data is available (adapted from Argent et al. (2008)).

This is not an advocacy against detailed spatial process descriptions. In depth knowledge of a specific (typically small scale occurring) process needs detailed descriptions. Also for larger scales, the incorporation of a more detailed description of a specific process can overcome a wrong or oversimplified conceptualisation. A good example is discussed by Arnaldos Orts et al. (2015), where the definition of the affinity parameter in the description of biochemical model kinetics is questioned (bacterial consumption of substrate, section 4.2). A central problem is the

conceptualisation of the mixing when using a lumped spatial domain, assuming it to be represented by the affinity parameter. A detailed hydrodynamic description explicitly describes the local conditions and mixing pattern by which the affinity parameter represents the specific function it is intended for in the model, i.e. the affinity of the bacteria towards the substrate. In fluid dynamics (air, water and soil) a detailed description is getting more feasible for larger model domains due to the increased computational power. Still, differences do exist between different media, with a soil matrix being much more heterogeneous than air.

In essence, the central message is that the modelling approach and the level of detail is a direct outcome of the available data, the objectives of the research and the characteristics of the modelled system. There is no reason to overkill the complexity of the model or assume a predefined structure, just because it is (technically) possible. The ability to easily experiment with different levels of complexity is of more importance than the quest for a universal model.

2.4.3 Protectionism towards the own creation

Quoting Andréassian et al. (2009) clarifies the point of protectionism with respect to hydrological modelling, but relevant in general:

> ... it sometimes seems as difficult for a hydrologist to publically admit the limitations of his creation as it is for an alcoholic to acknowledge his addiction.

This is put in a wider scientific setting by Chamberlin (1965), who warns about the parental affection towards the ruling theory causing to make facts fitting the theory and a tendency to find facts supporting the proposed theory. More recently, Nuzzo (2015) called this hypothesis myopia, making the researcher fixate on collecting evidence to support the hypothesis, while neglecting to look for evidence against it. Other explanations are not considered. This has, in most cases, nothing to do with fraud, but is caused by a cognitive bias that needs to be tackled.

One can easily see the analogy with the model developer searching for applications fitting the model representation and interpreting the simulations as *good fits* (whatever *good* may mean). As pointed out by Gupta et al. (2008) a lot of time and energy is still spent on attempts at model *validation*, in an attempt to defend the existing model, often without reference to any alternative model, hypothesis or theory. Beven et al. (2007) emphasize that by adopting existing *calibrated* models to only make *good* predictions it will be hard to learn about structural limitations.

Moreover, when model structure descriptions only point out the potential benefits of the model and do not clearly state the unavoidable assumptions, the choice of the most appropriate structure for any specific task is hampered (Todini, 2007). As a consequence, it is difficult for another user to judge the model suitability for another case study. If the assumptions and related weaknesses would be clearly communicated, it would give more guidance in the applicability and in possible improvements.

The lack in transparency in model descriptions from the model developer side is one thing, but the model user also bears the responsibility of making a proper selection of model structure. This selection is also driven by the availability of the code, easiness of use, the institutional settings and the experience of the user (Fenicia, 2008; Kavetski and Fenicia, 2011). Beven (2012) declared this as the natural tendency during model selection, to give a prior weight of one to his or her model and a prior weight of zero to all other models. This leads to applications of predefined model structures for specific systems that are questionable (Jakeman et al., 2006). Restricting yourself to a single model structure option is guiding the modelling study towards a too narrow direction, with the risk that there is no turning back. However, it is common practice to use already available, predefined one-size-fits-all model structures (Fenicia, 2008).

A good example of this *model-on-the-shelf* approach is illustrated by Herron et al. (2002). They state in their paper that the choice of models was governed by the clients' familiarity which *increased the acceptability of the results*. One could easily comment on this practice, however it provides at least an explanation for the decision of the specific models. In many other applications the same issue arises, but is just not reported. As mentioned by Buytaert et al. (2008), the success of the Soil and Water Assessment Tool (SWAT) model is partly due to the fact that it is freely available and not because it is in all these cases the most appropriate option. Taking into account the time-limited era in which research and consultancy need to be executed, this is perfectly understandable. However, it would be ignorant to not at least counter this with an improved scientific approach.

To be clear, this does not mean that these researchers are not considering the selection. In many cases, the decided model structure will be based on an expert-knowledge optimization, pragmatic modelling decisions and thoughtful evaluation of alternatives. However, questioning the structure of a model is something mostly performed in the initial stages of model development and considered as part of the research itself, and modellers thus rarely write about it. As such, approaches to questioning the structure of a model are more difficult to find and model failures are rarely fully reported in the peer-reviewed literature (Andréassian et al., 2012;

Kavetski and Fenicia, 2011; Sin et al., 2006). However, as stated by Beven et al. (2007), failures are also not reported due to the strong incentives to be positive and affirmative about the model, even if this results in predictions with models that not actually provided very good simulations. Moreover, when multiple alternatives may be considered when a model is developed, it is typical that only one approach is implemented and tested (Kavetski and Fenicia, 2011).

2.4.4 Monolithic and closed source implementations

In the past, comprehensive modelling systems have been constructed as large complex computer programmes (Beven et al., 2007; Buahin and Horsburgh, 2015). This is one of the reasons of the earlier described common practice of being restrictive to available predefined model structures. The monolithic implementation makes it difficult to adapt an existing model implementation, since it requires significant programming skills and time to revise the original source code, to understand the implementation and to adapt the algorithms for the required application (Buytaert et al., 2008).

Moreover, a lot of the source code used in environmental science is hidden behind license restrictions and commercial software applications. The latter is understandable, since this is one way of getting valorisation out of the scientific work and it is a useful way of bringing scientific outcome to a larger community of practitioners. However, when source code is not accessible, this inhibits the reproducibility of the results for the scientific community. The accessibility to model implementations has been pointed out earlier as a condition for reproducibility (Buytaert et al., 2008; Fenicia, 2008; Wilson et al., 2014).

Even when the implementation is available, modelling projects still can be difficult to audit and without a considerable effort, it is hardly possible to reconstruct, repeat and reproduce the modelling process and its results (Buytaert et al., 2008; Refsgaard, 2004). Fenicia (2008) rightly remarks that authors emphasizing the need for a flexible and modular approach (Beven, 2000), remain ignorant towards the application of fixed and monolithic structures, developed by their own (Beven and Kirkby, 1979).

The accessibility issue does not only appear to be relevant towards the model implementations itself, but also to the implementation of methods for model evaluation and analysis. Multiple methodologies for sensitivity analysis, uncertainty analysis and optimization are described in literature. These mathematical methods are often so complex that a full re-implementation of the computer code is

beyond the resources available to an environmental scientist (Buytaert et al., 2008). Besides, methodologies developed to optimise or estimate the predictive power of models are in many cases only reporting on a small set of applications, making it hard to evaluate the usability.

Even more important, the monolithic characteristic of model implementations limits the applicability of model comparison, since it obstructs the ability to **attribute** inter-model differences to specific processes and hypotheses.

Consider the following example first. When one would like to compare the quality of two types of tires for biking, it would not make any sense to put these two types on completely different bikes, cycled by two different cyclists on completely different types of roads. If then one tire deflates considerably more than the other, it could just as well be caused by any of the other circumstances and it would be wrong to attribute the increased amount of punctures to the quality of the tires. Putting both tires on the same bike (and regularly interchange them) or on two completely similar bikes (riding them under similar conditions) would be a far more effective strategy. Simply said, keep all the rest the same and only change the specific element targeting for.

The latter is however not possible with the monolithic model implementations that are regularly dealt with in environmental science. Model comparison studies to date have provided limited insight into the causes of differences in model behaviour, due to the impossibility of addressing the differences in modelled outcome to specific elements of each model (Clark et al., 2015b). When comparing models, there are often too many structural and implementation differences among them to meaningfully attribute the difference between any two models to specific individual components (Kavetski and Fenicia, 2011). In other words, when the performance of two monolithic (closed source) model structure implementations are compared, it is hard to know what exactly is causing a difference in performance. In that context, model comparison can only provide information about *better* performance, but systematic identification of model shortcomings is impossible.

2.4.5 Business as usual in model evaluation

The evaluation of the model structure is a continuous learning about the appropriateness of that model structure. Model calibration is an essential part of the evaluation (section 2.3). The inevitable mismatch between the researched system observations and the applied model output is partly compensated during the model calibration, making it a central element.

However, in many cases the evaluation is condensed to an optimization problem, instead of the exploration of the model performance from different points of view. Moreover, the optimization is performed using a single aggregated metric to quantify the difference between the model output and the observations. As such, optimization algorithms are applied to find the parameter set that minimizes the aggregated metric, ignoring the eventual lack of identifiability of the parameters. Some 'optimal' parameter set is determined by the optimization algorithm, but often the reliability of the estimate is not checked for.

In the validation step, i.e. the evaluation of the model output to an independent set of data, model performance is regularly reported in a very rough and simplified way (Gupta et al., 2008; Kavetski and Fenicia, 2011). Examples are the expression of model performance by statistics such as the Nash-Sutcliffe Efficiency (NSE) or a correlation coefficient, which does not directly test any individual hypothesis about the overall model (Uhlenbrook et al., 1999). It is recognized that such measures of average model output versus observations similarity lack the power to provide a meaningful comparative evaluation. The NSE summarizes model performance relative to the observed mean output, which is a very weak benchmark (Schaefli and Gupta, 2007). Nevertheless, the application is still frequently observed in literature.

The lack of information in the observations to discriminate between increasingly complex models leads to the acceptance of *equifinality* between models (Beven, 2006), meaning they are able to approximate the observations with equal performance. In some cases this will indeed be the conclusion based on the available observations. However, as pointed out by Gupta et al. (2008), if we have not properly tested the limits of agreement (or lack thereof) between our models and the data, this seems a lazy approach to science.

2.4.6 Intrinsic characteristics of environmental systems

Environmental systems are heterogeneous, open systems and modelling studies include a wide range of scales, as a conceptualisation of the processes involved. In a lab-environment some degree of control can be carried out, but once going to natural environments (e.g. catchments) it is very hard to control the experimental conditions and to identify clear system boundaries. Moreover, environmental systems are unique in their characteristics shaped by a specific geological activity, exposed to different climatological drivers and exposed to varying anthropological influences.

This *uniqueness of place* (Beven, 2012) of environmental systems is sub-field dependent and gets a lot of attention in hydrological applications where a catchment approach is predominant, but is equally relevant in water quality and ecological applications.

When compared to more controlled environments, such as in industrial chemical engineering, this uniqueness of place is less prevailing. The systems studied are according to a predefined design, active in controlled and closed reactors and easier to standardise. Hence, it is more convenient to propose a set of standard practices and guidelines among the scientific community. Models are still case dependent, but are mostly different configurations of defined system units that are reusable. The latter is done in flow sheet model software (GPROMS, 2015), which are nothing more than integrated software environments to couple different model units in order to mimic the case specific configuration studied. The development of the WEST platform for Waste Water Treatment Plant (WWTP) modelling (Claeys, 2008) is an example of the translation of the system units towards environmental science.

One could question to what extent this translation is possible to environmental systems, where the distinction between unit processes (e.g. hydrology versus hydraulics or sediment versus runoff) and the demarcation of each system boundary is far less clear. Hence, the uniqueness makes it much harder to come up with predefined system units that can be reused, which hampers collaborative progress.

The openness of environmental systems was already mentioned in section 2.3.2, where it was used as an argument against the possibility of validation. Any model will be falsified if we investigate it in sufficient detail and specify very high performance criteria. Even if a site-specific model is eventually accepted as *valid* for specific conditions, this still does not prove that the model is true (Refsgaard, 2004; Beven, 2012; Fenicia, 2008). This widely recognized problem of *uniqueness of place*, clearly illustrates that a modelling application should be site specific, being a function of the catchment characteristics, the data availability, and the modelling purposes. This highly contradicts the dominance of a few model structures in scientific literature and the monolithic implementations described earlier (Buytaert et al., 2008).

Dealing with an open, uncontrollable environment also induces limitations on the model evaluation. In disciplines such as physics, where the experimental conditions can be carefully controlled, it is often possible to rigorously apply concepts of statistical significance. In many environmental disciplines, events of interest may be infrequent or non repeatable, and the uncertainty in the observations is seldom

fully characterized (Kavetski and Fenicia, 2011). Moreover, experiments cannot be repeated under exactly the same boundary and initial conditions (Beven, 2000). For some environmental systems one has the luxury of optimal experimental design where inputs (such as to a bioreactor) can be manipulated to enhance the identifiability of a model. For most systems, however, we must at any given time accept the data that are available (Jakeman et al., 2006). Finally, models representing natural systems consist of multiple interacting components, making traditional (in a pure statistical manner) hypothesis testing less suitable and hindering the testing of individual modelling decisions (Bennett et al., 2013).

2.5 Overcoming conservatism: A model diagnostic approach

In the previous section some bottlenecks hampering the progress in environmental modelling were identified and listed based on existing literature. There are different initiatives already existing that aim to cure these conservative aspects of environmental modelling. Most of these comments are not new, but probably as old as the modelling practices itself and the scientific community is not ignorant towards the above mentioned pitfalls.

One could argue about the interconnection between the different bottlenecks raised in the previous section. The quest for a detailed all-in-one model description arose from the increased scientific insight. The growing technological possibilities lead to the creation of monolithic *all-in-one* model structures. Furthermore, the creation of monolithic models supports at the same time their usage by practitioners (people get trained to work with that specific model structure). However, that practice of model building brought emphasise on the model capabilities (a wealth of functionalities *what the model could do* rather than evaluation) giving rise to the curse of parental feeling towards the creation. The latter pushes attention to reusing the same model structure for new applications, leading to inferior practices of model structure evaluation and increased focus on model calibration (as in fitting parameters).

Independent from the accuracy of this statement (more an opinion than a hypothesis), it is clear that an alternative approach should be looked for.

Different authors have proposed an alternative model analysis to deal with the above. However, they differ in terminology, reasoning and historical framing. Based on the discussion in the previous section, overcoming the unidirectional

(pick model - fit to data - report -ready) usage of monolithic model structures, leading to the impossibility of testing individual modelling decisions, needs to be executed at three levels:

- Accepting the idea of **working hypotheses** and considering model structure building as an iterative learning process based on failures
- Making this model structure building practical and technically possible, with emphasis on **flexibility** in model development in an **open and transparent** manner, being a necessary condition
- Extending the scrutiny of **model structure evaluation** as being essential, moving beyond current model calibration practices

Furthermore, the different levels should be supported by a clear modelling terminology. In this section, the above three levels will be further clarified. The combination of these three elements will be referred in this dissertation as a **model diagnostic approach**, which provides a workable method and the conditions to counter the diagnostic problem definition provided by Gupta et al. (2008):

Definition 2.6. *Given a computational model of a system, together with a simulation of the systems behaviour which conflicts with the way the system is observed (or supposed) to behave, the diagnostic problem is to determine those components of the model, which when assumed to be functioning properly, will explain the discrepancy between the computed and observed system behaviour (adapted from Reiter, 1987).*

In other words, Gupta et al. (2008) considers the diagnostic problem as the search for deficiencies in a model structure. However, this definition of a diagnostic approach is too narrow, since it does not include the learning process and confines it to a single model structure. In this dissertation, the model diagnostic approach is defined in a broader sense, considering the necessity of flexibility in the model structure definition. The definition goes beyond the borders of the different communities within environmental modelling and supports a more common approach typically not encountered in literature.

2.5.1 Tier 1 of the model diagnostic approach: Multiple working hypotheses

As earlier described, the parental affection towards the own creation can lead to protective actioning. To guard against this, Chamberlin (1965) (which is actually

a reprint of the original article of 1890) urged for the method of multiple working hypotheses:

> The effort is to bring up into view every rational explanation of new phenomena, and to develop every tenable hypothesis respecting their cause and history. The investigator thus becomes the parent of a family of hypotheses: and, by his parental relation to all, he is forbidden to fasten his affections unduly upon any one.

The explicit consideration of alternative hypotheses within a transparent (i.e. reproducible) context is proposed in literature to guard against the cognitive bias towards scientific results (Nuzzo, 2015; Kavetski and Fenicia, 2011; Fenicia et al., 2014; Abramowitz, 2010; Beven, 2012). As stated by Beven (2012), this means that any model that predicts the variable of interest is a potentially useful predictor, until there is evidence to reject it. At the same time, this links the concept to the recognition of the principle of *falsification* of testable hypotheses well known in statistical testing (Popper, 1959; Kavetski and Fenicia, 2011). The latter basically means that one cannot accept a model, but that it can only be falsified and refuted. As such, it provides a response to the impossibility of model validation (section 2.3.2).

By accepting it, it places model development in a **rejectionist framework**, to detect what remains wrong about our conceptions of the model (Gupta et al., 2008; Beven, 2000, 2012). Or in other words, we can learn the most from model failures. Model deficiencies provide guidance about the potential improvements (Andréassian et al., 2012; Gupta et al., 2008). Andréassian et al. (2010) advocated that giving greater attention to the analysis of failures would be more beneficial for the advance of hydrological sciences (*Court of Miracles of Hydrology* workshop). The latter is actually relevant in all environmental sciences. Making this iterative learning curve explicit in the model development cycle, enables to communicate about these model failures in literature. Beven et al. (2007) refers the learning framework as a way to gear model structures to the specific conditions of each place (section 2.4.6). As such, the rejection of hypotheses for individual cases provides insight in the uniqueness of the place and the characteristics, referred to as *learning of places*.

Hence, we need to define some level of suitability, guided by the available observations (with observations in its broadest sense). When observations are incompatible with the model predictions, this suggests that the model can be rejected as a hypothesis of how the system works (Gupta et al., 2008). However, the rejectionist framework holds the possibility of accepting a poor model when it should be rejected (false positive or Type I error) or rejecting a good model when it should be

accepted (false negative or Type II error), as it is in statistical hypothesis testing. However, in contrary to statistical testing, the lack of replications of the observations in most environmental modelling cases, limits the applicability of statistical tests (Beven, 2012).

When observations are scarce and rejection of structure lacking, differentiation about the suitability needs to be made based on the observations at hand. A lack of differentiation leads to Type I errors. The testability of a model structure will increase in cases where an increasing number of output variables exists that can be compared to observations. As a consequence of the different uncertainties involved in modelling, any model can be rejected when sufficiently tested, leading to errors of Type II. The latter would be worse in a model diagnostic approach, since excluding good models would be a loss of information. Type I errors could hopefully be eliminated in the analysis by further evaluation (e.g. new data source) (Beven, 2012). For example, the notions of a flat or spherical earth were indistinguishable until new evidence was obtained by Magellan and others (Silberstein, 2006).

However, in many practical applications, no distinction can be made between the proposed set of model structures, due to an imbalance between the available observations and the model complexity. As such, this lack of differentiation, where none of the proposed model structures can be falsified with the available information (which can be interpreted both for different parameter sets of a single model structure as well as model structures), is also referred in literature as non-uniqueness, ill-defined, or, by philosophers, as underdetermination. Furthermore, within the hydrological community this is also referred to as **equifinality** as a generalisation of a lack of identifiability (Beven and Binley, 1992; Beven, 2000; Beven and Freer, 2001; Beven, 2006, 2008b, 2012). Notwithstanding the different contexts and interpretations, it refers to the same inability to differentiate (cfr. also distinguishability (Petersen, 2000)). In many cases, this problem of non-uniqueness is caused by a lack of *identifiability* of each of the individual model structures (section 2.3.3). As pointed out, the notion of identifiability is related to the possibility to give a unique value to each of the model parameters (Donckels, 2009), and the counterpart is also referred as overparameterisation. Hence, the influence and the interaction of the parameters is the key of the evaluation of model structures (focus of section 5.2).

The lack of identifiability leading to non-uniqueness, leads also to the *principle of parsimony* and parsimonious modelling (Wagener and Wheater, 2002; Obled et al., 2009; Young, 2003; Taylor et al., 2007; Willems, 2014) as a reaction on the quest for detailed model structures (section 2.4.2). The principle of parsimony, also stated as *Occam's razor*, is a problem-solving principle stating that among com-

peting hypotheses that predict equally well, the one with the fewest assumptions should be selected. It is in relation to falsifiability mentioned earlier, since simpler theories are better testable (Popper, 1959). In modelling terms, when choosing among models with equal explanatory power the simplest model is more likely to be correct. More degrees of freedom (i.e. parameters) makes the behaviour less dependent on the model structure itself (Kirchner, 2006). Hence, the latter increases the possibility of making Type I errors. On the other hand, a model structure that is too simple in terms of the number of processes represented can be unreliable outside the range of conditions on which it was calibrated (Wagener et al., 2001b). Overly simple models underestimate the prediction uncertainty when used to forecast outside the domain of the model identification (Reichert and Omlin, 1997). Combining the results of multiple models, each weighted by their respective likelihood, provides a practical solution to estimate the prediction uncertainty.

When aiming for parsimony, model structures should have the simplest parameterization that can be used to represent the observations (Wagener et al., 2001b; Sivapalan et al., 2003). The principle is also mentioned as the *dominant processes approach*, providing model structures that capture the key response modes of the system (Sivakumar, 2004, 2008). However, this principle of parsimony and the related terminology is embedded in the idea of identifiability analysis and directly follows the definition of practical identifiability analysis. If parameters are practically not identifiable, they do not comply with the idea of parsimony. So, identifiability is the preferred terminology, since it provides better the link with mathematical oriented literature dealing with this problem.

One could command that this learning process of multiple hypotheses does not fit within an engineering oriented modelling approach (section 2.3.2). The continuous rejection of model structures is in conflict with the necessity to create useful models for practical application. However, an engineering approach focusing on making predictions using a process-oriented model based on a conceptual representation (instead of a pure data-based model), is inherently making a hypotheses of the process descriptions. The difference is in the defined acceptance for suitability, which is a direct result of the proposed modelling objective. Whereas in a scientific oriented approach the aim is understanding and the search for model deficiencies is central, leading to a high level of rejection, the engineering approach is defining suitability purely in the purpose of providing reliable predictions. Both approaches rely on a case specific approach and learning curve, the difference is in the scrutiny of evaluation. Nevertheless, both need to have a sufficient set of tools to perform

the required model evaluation in which the identifiability of the parameters is crucial to evaluate the competing hypotheses.

2.5.2 Tier 2 of the model diagnostic approach: Flexible model development

An essential consequence of the model application of monolithic model structures is the impossibility to properly compare different model structures and to address model failure to specific modelling decisions (Kavetski and Fenicia, 2011) (section 2.4.4). The key purpose is to isolate and evaluate individual processes and modelling decisions as much as possible. When the number of differences between alternative model structures is kept to a minimum, it is possible to **attribute** the differences towards the individual modelling decisions. Under these conditions, the previously described multiple hypotheses approach becomes meaningful (Clark et al., 2015b).

The aim is to enable a controlled and systematic evaluation both on overall model structure as well as on individual components. By focusing on the level of model subcomponents representing individual processes, it becomes possible to select the best components from different models and as such, to avoid the need of rejecting entire models. Hence, individual modelling decisions are actually nested together into modelling decisions on a higher level (i.e. hierarchical level). As an example, the decision to take into account the degradation of a specific chemical component by bacteria is one hierarchical level higher to the decision of the specific kinetic (e.g. Monod) this process is described by. As such, this hierarchy is an important characteristic in the structure development. By scaling to coarser levels, this approach directly fits into an integrated modelling framework, where interchanging of model components is possible on the different hierarchical levels.

Besides the ability to attribute modelling decisions, the highly heterogeneous properties of environmental systems require a very tailored and specific approach to the representation of a system. Model applications are always case specific. They are a function of the system characteristics (e.g. the boundaries to demarcate the system), the specific modelling purpose and the available data (Fenicia, 2008). From this, flexibility appears to be a logical design criterion for modelling to suit local conditions (Beven, 2008b). In a recent review focusing on hydrological modelling in urbanized catchments, Salvadore et al. (2015) regard flexibility in terms of spatial and temporal discretization, model components and input requirements as the key characteristics to handle the huge diversity of situations.

However, instead of flexibility, the modeller is regularly faced by a choice between existing model software, each providing limited flexibility, i.e. only permitting the adjustment of parameter values. As mentioned by Leavesley et al. (2002), model development should shift from the question *which is the most appropriate model structure?* towards *what combination of process conceptualisations is most appropriate?*. Making this process selection transparent is key for proper revision of the suitability of the selected structure and a first step towards supporting practitioners of doing so.

Not all flexible modelling environments support the requirements to enable the model development approach as presented. In their publication about pursuing the method of multiple working hypotheses and focusing on hydrological modelling, Kavetski and Fenicia (2011) proposed the following key requirements for flexible modelling frameworks:

1. Support **multiple alternative decisions regarding process selection and representation**, which means that the framework should provide multiple options for describing individual processes, e.g. the representation of different kinetics to describe conversions or the representation of interception by vegetation.

2. Accommodate **different options for the model architecture**, representing the connectivity between different model components. Here the focus is both on variation in which processes to combine, as well as on the spatial configurations.

3. The ability to separate the hypothesized model equations from their solutions, especially if the latter require numerical approximations. In other words, **the mathematical and computational model should be clearly defined and separately identifiable**. This is particularly relevant for hydrological modelling, where the division between the model equations and the numerical implementation is often lacking and not communicated (Clark and Kavetski, 2010).

As pointed out by Buytaert et al. (2008), a central point is that **model codes should be fully accessible, modular and portable**. In order to adapt individual model elements on all hierarchical levels, the possibility to change source code is a necessary condition. Hence, the requirement of **accessibility of the source code** on the process level is an important requirement not stated by Kavetski and Fenicia (2011). Providing readable source code and proper documentation are important as well, although not a necessary condition to test individual modelling decisions and rather general good scientific practice.

Flexible modelling environments

It is understandable from a historical perspective that the development of model implementations was done as a monolithic unity. However, frameworks have always been developed to provide the ability of building alternative model representations, with varying level of granularity. In essence, any modelling framework that enables experimenting with different ways of representing the system, supports the multiple hypotheses approach (Kavetski and Fenicia, 2011). Hence, the antidote for monolithic modelling can be referred as component-based modelling, modular modelling or loose model coupling (Buahin and Horsburgh, 2015; Claeys, 2008). It involves decomposing a complex system into smaller functional units called *components* that have specified interfaces, which allows them to be coupled together to represent a larger and more complex system. The set of components can be coupled in a hierarchical manner to form complex systems, also referred to as hierarchical system modelling (Filippi and Bisgambiglia, 2004).

Within the scope of integrated environmental modelling, the creation of modular frameworks is well-established (Argent et al., 2006; Filippi and Bisgambiglia, 2004; Krause et al., 2005; Bach et al., 2014; Laniak et al., 2013; David et al., 2013). Modular modelling approaches allow creating environmental models from basic components (Argent, 2004, 2005), which makes composing model structures less time-intensive and which can be applied within the scope of a webservice based technology (Vitolo et al., 2015). Recent developments are capable of dealing with both spatial and temporal misalignment in between the coupled components, i.e. the individual components operate on different spatial resolutions and time steps (Schmitz et al., 2014).

Many of these integrated modelling environments are in essence generally applicable and independent of the scientific application. Still, most of them are case and discipline dependent (Argent, 2005). Hence, a huge set of environments, software and standards do exist: focused on hydrological/hydraulic modelling (Leavesley et al., 2002; Clark et al., 2008; Wagener et al., 2001a; Bach et al., 2014; Welsh et al., 2013), ecosystem and ecological modelling (Voinov et al., 2004; Villa, 2001), water quality and waste water simulation (Reichert, 1994; Vanhooren et al., 2003; Claeys, 2008), chemical and industrial applications (flowsheet simulators) (GPROMS, 2015), earth systems modelling (Peckham, 2008; David et al., 2013) and general spatial models (Argent, 2005; Wesseling et al., 1996). The construction of these models can be done with an explicit coupling framework connecting components in a user interface (Vanhooren et al., 2003) or by a provided coupling standard such as the open-MI standard (Gregersen et al., 2007), which increases user accessibility and prevents new implementation. It is also done

by a model language approach (Wesseling et al., 1996; Kraft et al., 2010), where functions and building blocks are represented by coded definitions. The latter approach of using scripting tools got the advantage of being easily extended and at the same time it can be used as a 'glue' to external models or components (Kraft et al., 2010).

Spatially distributed models are taking advantage of spatially distributed forcing and process descriptions to describe the system (Tang et al., 2007a). Spatial development of flexible model structures needs a computational system that couples and coordinates modules in a simulation together with a Geographic Information System (GIS) to perform the spatial analysis within the simulation environment. Both user interface (Pullar, 2004; Changming et al., 2008; Maxwell and Costanza, 1997) and model language approaches (Fall and Fall, 2001; Wesseling et al., 1996) do exist. Wesseling et al. (1996) developed the dynamical modelling language PCRaster, which can be used to construct spatio-temporal models and can be called from the Python programming language.

However, many of these modelling frameworks provide flexibility on a coarse grain granularity, which does not allow to isolate and investigate individual modelling decisions (Clark et al., 2015b). Moreover, coupling of model structures can cause the models to interact badly (Abramowitz, 2010; Voinov and Shugart, 2013). To maximize the possibility for hypothesis testing, modular modelling frameworks should be accessible on a finer granularity (Clark et al., 2011a).

A more in depth literature study of all existing frameworks and confronting them with the requirements presented above is out of scope of this dissertation since it should be regarded from a software development point of view as well. Moreover, the lack of transparency in many of them would hinder such an analysis. Still, it can be summarized that many of these modelling frameworks do rely on Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs) as the underlying mathematical structure, i.e. a continuous dynamical process description. Components are mostly entities defined for a specific domain (and its boundaries) for which a balance is defined (mass, momentum, energy) and for which processes need to be assumed that define the incoming, outgoing and conversion terms.

As such, for convenience this dissertation will focus on fine grain level variations that are directly enabled by the implementation of ODE models represented by Equation 2.1. This directly complies with the requirements since it enables complete flexibility in the process representation and architecture. Moreover, the source code (python programming language) is directly available, since it is not dependent on any existing software environment enlisted earlier. Moreover, some

of the existing software supporting flexible modelling, such as Aquasim (Reichert, 1994) and West (Claeys, 2008) (amongst many others), support direct implementation of a set of ODEs.

2.5.3 Tier 3 of the model diagnostic approach: Extended model evaluation

It is clear that a single performance metric will not provide a sufficient basis for characterizing all relevant aspects of model performance, let alone the possibility to differentiate the suitability of different model structures or identify deficiencies on the process level (section 2.4.5). Aggregated metrics of model performance lack the ability to distinguish between individual modelling decisions because of the interaction between model components (Kavetski and Fenicia, 2011). Finding model failures is in practice not always straight-forward and requires a more in depth evaluation by using extra data sources (Anderton et al., 2002), a combination of multiple metrics (Gupta et al., 2012) or model evaluation tools (Bennett et al., 2013). As such, the recognition of multiple working hypotheses must be combined with the development and application of stringent model diagnostics that challenge both individual constituent hypotheses and the overall model structure (Kavetski and Fenicia, 2011).

For practical application, this can be achieved by the ability to get as much out of the available observations as possible, i.e. maximize the amount of information that can be extracted from observations. On the other hand, it comes down to the ability to check model behaviour in as many different ways as possible. One can look into the model itself or in confrontation with observations. Uncertainties arising from both the model structure and the observations are inherently present. They will limit the ability of the analysis and a proper attribution of the involved uncertainties is essential. In other words, the observational data provide (imperfect) evidence regarding the true state of the system (Bennett et al., 2013).

Model evaluation is directly linked with model calibration, since the adjustment of model parameters is steered by the performance of the model to observations, aiming to maximize performance. However, in the diagnostic approach aimed for, the focus moves from model calibration as a parameter adjustment exercise towards model structure evaluation as a combined process of component and parameter adjustment, giving them equal importance. The latter is also supported by the (sometimes) subjective and imprecise distinction between the model structure (the model equations) and the model parameters (the adjustable coefficients in the model equations) (Kavetski and Fenicia, 2011). As a straightforward example to

clarify this, take the sameness of the *structurally different* equations $y = k \cdot x^{\alpha}$ and $y = k \cdot x$ when the parameter $\alpha = 1$. More general, algebraic expressions and *different* systems of differential equations can behave functionally very similarly depending on the range of application and parameter values.

The ability of translating the process of model evaluation into a fixed recipe style workflow is probably close to non-existing (Fenicia (2008) refers to *the art of modelling*). However, to support an appropriate diagnosis of a model structure, it is essential to develop a set of tools that can be applied and combined in as many ways as possible. Just as a medical doctor examines a patient using different technologies (from stethoscope to X-rays) to identify failures, a modeller needs to have a range of tools to identify model deficiencies, going from quick visual exploration to computer intensive algorithms.

2.6 Conclusion

This chapter provided an introduction to some general concepts of modelling and some clarification on the incoherent terminology encountered in the scientific literature. Besides the problem of terminology, other bottlenecks are identified that are currently hampering the development of improved modelling practices and lead to conservative practices. These factors are relevant for different environmental modelling disciplines, making it useful for a wide audience.

From the identification of these bottlenecks, a diagnostic approach with three main tiers is defined, which tries to counteract these bottlenecks in a structured way. To fulfil the aim of a generic framework, these tiers are providing general boundary conditions and requirements: the acceptance of multiple working hypotheses, the necessity of flexibility in the model development inherently linked with the minimal requirements on the technical implementation and the necessity of a shift from current parameter adjustment practices towards the evaluation of individual model decisions. The necessity of an open and transparent implementation of models is generally ignored in literature, but appears to be an essential condition to overcome the existing conservatism.

PART II

MODEL DIAGNOSTIC TOOLS

CHAPTER 3

Model structure diagnostic tools

3.1 Introduction

The uniqueness of place, the available data and the research questions involved within the scope of each individual environmental model study require a tailor-made approach in model construction. Similar to the necessary flexibility in model construction, the model structure evaluation as well needs to be adaptive towards this intrinsic variability in modelling studies. Strategies to diagnose model structures in terms of performance, uncertainty, identifiability and complexity are necessary.

The lack of parameter identifiability is an important indicator to diagnose model structures (see chapter 2). A lack of parameter identification results in the incapability of finding an identifiable set of parameters for a specific model structure as well as the incapability of finding a model structure outperforming other model structures. Hence, methods to evaluate parameter sensitivity and identifiability are an essential element of model structure evaluation.

However, many scientific papers start from a predefined theoretical (statistical, possibilistic...) framework and its underlying assumptions, although it is not always clear (and mostly not even discussed) if the chosen framework is the most suitable one for the application at hand. Hence, a similar problem of being preconditioned about the used methodology occurs as is the case for model structure selection, leading to business as usual in model evaluation (section 2.4.5). However, classical fitting methods lack the power to detect and pinpoint deficiencies in the model structure. These methods assume that the residuals (i.e. the difference

between the observations and the model output) behave statistically similar as the error of the observations (uncorrelated, with zero mean and uncorrelated) and assumes that the model structure is correct, which is in many cases not justifiable (Vrugt and Sadegh, 2013).

From a model structure selection point of view, the main idea is to get insight in the model structure characteristics and behaviour in order to assess its suitability. A rigid theoretical framework can be useful, but not before sufficient insight is gained about the model structure properties. Just as it is essential in statistical data analysis to first perform a data exploration under scrutiny before applying any statistical model, one should first diagnose the model structure with respect to the provided data and specific research objective in as many ways as possible. In a next phase, a more theoretical framework can be applied.

Consider the following simplified real-world case to illustrate this: modeller X is interested in predicting the transgression of the ammonia concentration limits defined by the regulation in a river segment. Too often, current sensitivity analysis applied for these models assesses parameter sensitivity on the average of the simulated concentrations in time. However, the research question focuses on exceeding a concentration limit, so the user should be able to easily check parameter sensitivity towards maximum concentration values, concentrations above the threshold and the effect on false positive or negative trespassing of the regulation value. Moreover, in function of model calibration, these same metrics will be useful to evaluate model performance. Understanding, availability and easy of development and application of these metrics is therefore essential to enable the analysis.

Other research questions will ask for alternative metrics. As such, the ability to easily apply a variety of these metrics in combination with a wide range of (existing) model evaluation methods is the focus of this chapter. By doing so, the aim is to overcome the conservatism in model evaluation as denounced by Gupta et al. (2008) and Kavetski and Fenicia (2011) which can be achieved by empowering both scientists and practitioners in the exploration of model structures. This exploration is mainly driven by the research question and needs to be supported by a wide set of tools that are easily applicable. In a later stage, this can converge to a decision about a theoretical framework (e.g. least square estimation, formal likelihood definition...).

This chapter provides a broad overview of existing methodologies, from the point of view of defining a proper set of performance metrics. As such, it attempts to provides a pragmatic and practical answer towards the development of a more robust method of model evaluation proposed by Gupta et al. (2008). Their diagnostic approach describes the usage of signature behaviours and patterns observed

in the input-output data to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be improved for the purpose of learning and scientific discovery. Furthermore, it anticipates to the request of Bennett et al. (2013) for a more generalised repository of evaluation approaches across the spectrum of environmental modelling communities. In contrast with earlier work of Moriasi et al. (2007), the aim is not to provide a fixed step by step scheme for model evaluation.

The chapter is structured as follows. First, the central position that metrics have in model evaluation algorithms, will be discussed. Next, the construction of aggregated and performance metrics will be described. Due to the dependence of many algorithms on numerical approximation by sampling techniques, the chapter finishes with a general introduction on sampling techniques, as used by a large number of existing methodologies.

3.2 A plethora of frameworks

Similar to the monolithic characteristic of model structure implementations (section 2.4.4), an excessive focus goes to the usage and communication of (apparently) different model evaluation methodologies, each provided with a unique acronym (DREAM, IBUNE, BATEA, NSGA, Parasol...). This leads to an overflow of potential options on the one hand, but to a general conservatism in the application on the other hand (section 2.4.5). Powerful algorithms are being developed, capable of handling high-dimensional parameter spaces of non-linear models (Vrugt, 2015). However, methodologies for environmental model evaluation look more like a *bunch of tricks*, due to a lack of integration, rather than a consistent scientific discipline. Environmental modellers face the existence of a wide range of useful, but highly repetitive, non interoperable model evaluation techniques (Matott et al., 2009).

The overview of software based tools (65 different model evaluation tools) enlisted by Matott et al. (2009) illustrates the wide variety of existing methods and options. However, many of these methodologies are using the same building blocks and Matott et al. (2009) detected a considerable amount of overlapping functionality in the assembled list of tools. Methods with different names or developed for different applications are sometimes more similar than they at first sight appear to be (Bennett et al., 2013). By dismantling existing methods and pulling apart the algorithm from the supporting modules, many similarities can be identified and as such, reused when implemented in a more modular design.

3.2.1 Features of evaluation methodologies

A complete unravelling of all the existing theoretical methods as well as their implementations would be infeasible, due to the huge variety in described methods in literature. However, some typical characteristics of environmental modelling results in a regularly seen pattern.

The monolithic and closed source properties of model implementations (section 2) gave rise to evaluation methods that can communicate independently of the model structure itself. This explains for example the limited usage of structural identifiability methods such as the Laplace transform method or the Taylor series method which require direct interaction with the model equations (Dochain and Vanrolleghem, 2001).

In commonly used methods, communication is performed by simulating the model with different model inputs (e.g. initial conditions, parameters...) and extracting information from the available model output state variables without direct handling of the differential equations itself. Moreover, the non-linear behaviour of the underlying mathematical equations leads to the impossibility of finding an analytical solution when working with most of these (probabilistic) methods, explaining the popularity of Monte Carlo (MC) and other numerical approaches. The non-linear properties of environmental models also steered the development towards global methods, i.e. inspecting the whole parameter space instead of the direct neighbourhood of a single parameter combination. Model non-linearities induce more complicated shapes of the objective parameter response surface with multiple optima. Global methods reduce the risk of getting trapped into local optima of the parameter space (Nopens, 2010).

Extensive work has been performed earlier on local methods (Dochain and Vanrolleghem, 2001; Reichert and Vanrolleghem, 2001) and the capabilities of an iterative application of local methods (also referred as *robust* approach) is essential to mention here (Rodriguez-Fernandez et al., 2006; Donckels, 2009; De Pauw, 2005). Development of robust methods is ongoing parallel to this dissertation, with the application of new methodologies (Van Daele et al., 2015a) for which the implementations are made available (Van Daele et al., 2015c).

As such, the focus in this chapter (and dissertation) will be on methods that work **independently** from the model implementation, that can be **approximated numerically** and that are screening the **entire parameter space**.

3.2.2 A metric oriented approach

Let us first assume that we would have an infinite number of model simulations available for a specific model application. By doing so, the discussion about sampling strategies and the curse of dimensionality can be ignored at this point (the necessity of numerical approaches will be explained later). The advantage of this (non-realistic) assumption is that it supports the idea of an exploratory model structure handling. It removes the obscurity caused by the current focus on sampling strategies. Moreover, it avoids at the same time the blended discussion with optimization algorithms to find an optimum. The most optimal set (no matter on how this is defined) is already a subset of the entire set of available simulations. The optimization boils down to the idea of deriving the min or max of a specified model output derived metric (e.g. Sum of Squared Errors (SSE)). As such, an understandable and easy workflow pipeline can be constructed amongst many of the regularly used methods in literature (Figure 3.1). Similar to Gupta et al. (2008), it puts the focus on the **calculation of output-derived aggregation metrics and performance metrics**. Although this is a rather trivial consideration, the necessity to emphasise on this aspect is augmented by the conservatism in current model evaluations, neglecting or simply ignoring the selection of evaluation metrics and, for simplicity, just running the plethora of methods with *default* and *predefined* settings (Gupta et al., 2008).

As such, a model structure evaluation starts with the translation of the research purpose into a set of model evaluation metrics (section 3.3 and section 3.4). Based on the selected metrics, different techniques can be selected to visualize and interpret the model structure characteristics. In our utopian case of an infinite availability of simulations, possibilities are countless and can be easily translated into well-known modelling concepts:

- Finding the optimal (lowest/highest) value for a wide range of metrics: single and multi-objective optimization
- Assessing and visualising the effect of changing parameter values on model outputs to discriminate parameters with large impact from minor influencing parameters: sensitivity analysis
- Evaluating the sensitivity of the output to the parameters in combination with the parameter interaction: identifiability analysis (cfr. section 2.3.3)
- Exploring the posterior distribution conditioned by a set of observations: parameter uncertainty estimation

Most of these operations are a matter of correctly selecting a subset of simulations to provide a visual or quantitative summarizing evaluation of the analysis. As mentioned earlier, optimization is a min or max operation, sensitivity analysis can start from the scatter plot of the aggregated model simulations in function of the parameter values and sensitivity indices (such as Sobol) can be derived from this. Based on a scatter matrix plot of the parameter values with a colormap addressing the performance metrics, insight in parameter interactions and posterior parameter distributions can be derived graphically. Hence, for lower dimensional applications where current computational power is able to mimic the situation of unlimited model simulations, these graphical methods should be addressed within a first exploratory analysis, equivalent to the descriptive part of a statistical data analysis.

When we take the reality of environmental models into account, i.e. working with high-dimensional parameter spaces and non-linear equations, a more efficient parameter space exploration is needed and the results of these exploratory eval-

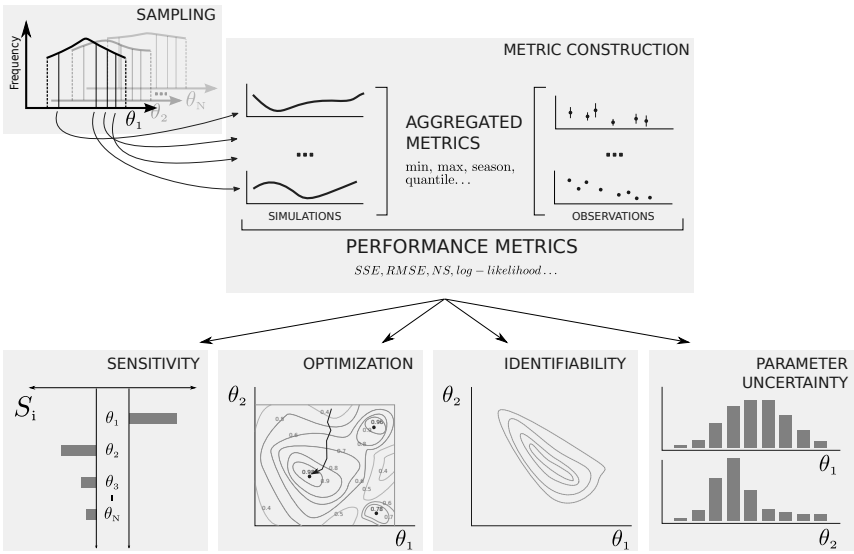


Figure 3.1: Simplified representation of the main model evaluation methodologies, neglecting theoretical assumptions and assuming an infinite ability to run simulations. A multi-variate parameter (input) space is sampled to generate a large set of simulation outputs which can be recalculated to derive aggregated metric values, compared to observations in any performance metric or any combination of these two. These metrics can be used by a variety of methods for sensitivity analysis, uncertainty analysis, identifiability analysis and optimization.

uations should be treated carefully. For example, the aim of optimization is to find the optimal value as fast as possible by iterating through the shortest path. The development of Markov Chain Monte Carlo (MCMC) approaches to sample the posterior probability density function of the parameters is another well-known example. However, from this perspective, the MCMC is not the purpose on itself, but a sampling method to increase the efficiency in approximating a posterior density function (section 3.5).

From this range of model evaluation techniques, the application of sensitivity analysis is of particular interest for model structure evaluation (Wagener and Kollat, 2007). It also provides support in the prioritization of important parameters, insight in the model structure characteristics and identifiability by revealing parameter interactions (introduced in section 2.3.3). Hence, these methods are of major interest and will be explained in more detail in chapter 5.

A clear communication about the difference in elements focusing on environmental model behaviour (domain-specific metrics) and elements aiming for an increased efficiency (general) is crucial. The latter is a problem exceeding the borders of the environmental modelling community, which is something that should be taken advantage of by trusting available libraries and packages from other disciplines that provide these functionalities. In practice, the availability of a callable function that calculates the metric as function of the model inputs (in many cases the parameters) is an essential step to enable the coupling with a wide range of existing libraries and packages.

Separation of the sampling (dimensionality) problem (section 4.1), the construction of the required performance metrics (section 3.4) and the analysis itself is crucial and should be reflected in the implementation architecture as well. To ensure the building blocks for each part are reusable to the at most extent, modularity is a key element to anticipate for this. For the environmental modeller, the ability to easily create and use different metrics when applying a model evaluation methodology, is essential as it provides the link between the research question and the set of available algorithms. The creation of these metrics will be further dealt within the following sections.

3.3 The construction of aggregated model output metrics

The computation and checking of aggregated model output metrics practically boils down to handling observed and modelled time series of all kind. However, the lack of standardisation, the huge variability in output formats and data type descriptions and the regular absence of proper meta data hampers this stage of the work, partly resulting into the conservatism in model evaluation application (cfr. section 2.4.5). Still, this time-intensive part of the research is generally ignored in scientific communication.

Nevertheless, environmental scientists are dealing with observation records frequently. Reading in time series, transforming them and extracting specific periods for visualisation and analysis are part of the daily work. The derivation of aggregated metrics for model evaluation is just another type of time series manipulation.

It is noteworthy that these aggregation metrics can be seen as part of the general model definition provided in Equation 2.1. The algebraic part of the model definition defines a set of functions $\boldsymbol{g}(\mathbf{x}(t), \mathbf{y}_{\mathrm{t,in}}(t), \boldsymbol{\theta}, t)$ that maps the time dependent state variables $\mathbf{x}$ of the model into the variables of interest $\hat{\boldsymbol{y}}$. This can be just a selection function (subset of state variables), but it also can be a wide range of aggregation functions. Hence, the aggregation functions applied can be interpreted as part of the model itself and communicated as such. At the same time, it should be noted that performance metrics are not part of the model definition, since they require some kind of observations to compare with.

To support the community of environmental engineers, the development of easy to use tools to perform this kind of aggregations in a documented and automated way, are essential. In order to link the calculation of the aggregation with the range of existing algorithms, it needs to be callable as a function by these algorithms. Spreadsheet software is still regularly used to calculate aggregated metrics, but does not easily support an automated operation as callable function. The lack of automation and inherent documentation of the calculation steps results in repetitive work when dealing with large amounts of data (Markowetz, 2015). Scripting languages like R (R Core Development Team, 2008) and Python (Rossum, 1995) on the other hand, provide flexibility, enable automation and reproducibility and increase efficiency.

The necessity of tools to facilitate this kind of aggregations is generally not discussed in literature, but the execution can require a substantial amount of time. Therefore, the availability of tools that automate these aggregations can support practitioners in extending model evaluation. Within the scope of this dissertation, the hydropy Python Package 1 has been created, relying on the power of Pandas (McKinney, 2010), a powerful environment for data analysis. The added value of the hydropy Python Package 1 is a set of implemented functionalities to select specific parts of a hydrograph.

Python Package 1 (hydropy)**.**
The hydropy package supports the fast handling and selection of time series records, with a domain specificity towards hydrological applications. The package originated from making the routines performed in this dissertation reproducible and from making the functionalities available, created within the scope of the project performed by Van Hoey et al. (2014a).
The package adds a layer of domain-specific functionalities on top of the existing Python Pandas package (McKinney, 2010). As such, the power of Pandas is enabled, while focusing on a domain specific set of functionalities (Van Hoey et al., 2015a).

(https://stijnvanhoey.github.io/hydropy/)

3.4 Construction of performance metrics

Quantitative performance metrics compute the difference between model output and observations and are essential to evaluate model performance. These metrics are used to translate a model calibration into an optimization problem as well as to provide quantitative information about model performance. Here, we will use the term performance metric as a generic name for any quantitative metric or function used to evaluate model performance (also referred to as objective functions) or to condition parameter values (e.g. likelihood functions). It depends on the chosen metric and the related assumptions to what theoretical framework one is subjected, Ordinary Least Squares (OLS) is probably the most known. The assumptions linked to OLS do restrict the user, but at the same time OLS provides additional features. For example, the usage of a least squares approach provides a convenient derivation and communication about parameter confidence intervals.

Still, similar to any theoretical framework, these assumptions need to be verified for their validity.

Hence, existing theoretical methodologies proposed in literature can also be interpreted as alternative expressions of the search for diagnostic measures, i.e. performance metrics (Gupta et al., 1998). However, in terms of a model diagnostic approach (which goes beyond finding an optimal parameter set), a single aggregated performance metric is mostly not sufficient in evaluating the performance, since it lumps time-dependent information. Hence, within the learning process of model evaluation, flexibility in the usage of different metrics is essential and the applicability of this range of metrics for sensitivity analysis, uncertainty analysis and optimization algorithms is crucial. This is commonly ignored in literature, resulting in a limited set of performance metrics reused over and over again. The performance metric construction will be further elaborated on in this section.

3.4.1 Classification of performance metrics

Some performance metrics are frequently used in literature and are either applied during the model calibration or as post processing evaluation of the estimated model fit. Many alternatives can be applied and existing papers provide an extended set of performance metrics (Gupta et al., 1998; Legates and McCabe Jr, 1999; Moriasi et al., 2007; Dawson et al., 2007; Gupta et al., 2009; Hauduc et al., 2015; Pfannerstill et al., 2014).

Central in the construction of performance metrics are the residuals, which is the difference between the modelled output values $\hat{\mathbf{y}}$ and the observed values $\mathbf{y}$. The modelled values can be any aggregated metric from a variable of interest (section 3.3), as long as there is a corresponding observed value (either by direct measurement or by using an aggregation function on the measurements as well).

Hauduc et al. (2015) classify performance metrics in following main classes:

- **Event statistics**: When the accuracy of specific events is required, such as storm flows or toxic peaks, the performance during these specific events needs to be evaluated. Examples are the peak difference (Gupta et al., 1998) or the difference in timing of the peak values.
- **Absolute criteria from residuals**: The absolute criteria are based on the sum of the residuals (which can be raised to a power), generally averaged by the number of observations available. Low values suggest good agreement.

Examples are the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad ; \quad MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{3.1}$$

- **Criteria evaluating event dynamics**: These metrics penalize noisy time series and model outputs with a timing error, such as the timing of a peak value. As an example, the Mean Squared Derivative Error (MSDE) is the square of the differences of predicted and observed variations between two time steps:

$$MSDE = \frac{1}{N-1}\sum_{i=2}^{N}\left((\hat{y}_i - \hat{y_{i-1}}) - (y_i - y_{i-1})\right)^2 \tag{3.2}$$

- **Residuals normalized with observed values**: For these metrics, the residuals of each individual measurement are divided by the observed values itself, balancing the effect of large errors related to large values of the variable. Low values suggest good agreement. An example is the Mean Square Relative Error (MSRE):

$$MSRE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2 \tag{3.3}$$

- **Sum of residuals normalized with sum of observed values**: Instead of dividing the individual residuals by their corresponding observations, the division is performed on the entire set of observations, such as with the Percent Bias (PB):

$$PB = 100 \cdot \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{\sum_{i=1}^{N}(y_i)} \tag{3.4}$$

- **Comparison of residuals with reference values and with other models**: These performance metrics compare the residuals with residuals obtained with any reference model, for which the mean value ($\bar{y}_i$) is most well-known as defined by the regularly used NSE metric:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{3.5}$$

Other classifications do exist to classify existing metrics (Dawson et al., 2007). The difference between the focus on low and high flow can be compensated by

the transformation of the observations and modelled values with a Box-Cox or logarithmic transformation (Thiemann et al., 2001; Willems, 2009) (section 6.3.3). Furthermore, the transformation can be useful to support the homoscedasticity of the residuals, enabling the applicability of a theoretical (probabilistic) framework as applied in Dams et al. (2014).

The classification provides a first guideline to the combination of performance metrics, as combining metrics of different classes will be more effective as compared to using multiple metrics within one class.

3.4.2 Metrics as estimators

In the previous section, the difference between the modelled output and observed variable was described as any kind of performance metric quantitatively. However, many of these performance metrics can be seen as special cases of Maximum Likelihood (ML) estimators under specific assumptions. Some important concepts will be introduced here according to the approach of Dochain and Vanrolleghem (2001). Further excellent reading material is also provided by Reichert (2003) and MacKay (2002).

Maximum likelihood Estimation

The ML approach is a fundamental part of a frequentistic statistical approach for defining statistical estimators. It enables the estimation of the parameters of a statistical model given a set of observations. Considering the potential values the parameters of the model structure can have, it is intuitive to think that some values are more likely than others. More likely values will correspond to a smaller difference between the modelled output and the observed values.

A classical frequentistic approach starts from the idea that the available observed values are a sample of the universe of observations (i.e. considered as random variables $\boldsymbol{Y}$). These frequencies can be expressed as a density function $f(\mathbf{y}, \boldsymbol{\theta})$ (statistical model), which depends on a set of parameters $\boldsymbol{\theta}$. The Probability Density Function (PDF) $f(\mathbf{y}, \boldsymbol{\theta})$ within a process based model approach, describes both the deterministic model (environmental model structure) and a stochastic part (Omlin and Reichert, 1999). In that case, the statistical model describes the (measurement) error term around the values of the deterministic process based model, assuming the model structure to be correct.

The available observations are considered as realisations of the random variables $\boldsymbol{Y}$. The estimator $\hat{\boldsymbol{\theta}}$ is function of these random variables and is therefore a random variable itself, in contrast to the (real, but unknown) model parameters $\boldsymbol{\theta}$. Further discussion about the properties of estimators is out of scope here, but it is important to understand that the maximum likelihood approach is a general method to find such an estimator $\hat{\boldsymbol{\theta}}$.

The PDF $f(\mathbf{y}, \boldsymbol{\theta})$ gives the probability density for observing the values $\boldsymbol{y}$ given the parameters $\boldsymbol{\theta}$. The probability distribution of the observations given the parameters is expressed as $P(\mathbf{y} \mid \boldsymbol{\theta})$. So, the ML approach identifies the setting of the parameter vector $\hat{\boldsymbol{\theta}}$ that maximizes this probability $P(\mathbf{y} \mid \boldsymbol{\theta})$ for the available set of observations, leading to an optimization problem. By combining (multiplying) the individual probabilities, a likelihood function can be constructed for a given application. Notice the mixed usage of probability and likelihood. Actually, the likelihood is defined as the probability of the observations as a function of $\boldsymbol{\theta}$.

Ordinary Least Squares Estimation

OLS is a special case of ML estimation. When the assumption is made that the model is represented by independent (uncorrelated) errors originating from normal distributions, the likelihood function is provided by the product of the individual normal distributions:

$$L(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2}} \exp\left[-\frac{1}{2}\left(\frac{\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{i}}}{\sigma_{\mathrm{i}}}\right)^2\right] \tag{3.6}$$

with $\hat{\mathbf{y}}$ the model variables which are a function of $\boldsymbol{\theta}$ (in this case all part of the process based model), $\mathbf{y}$ the set of observations and σ_{i} is the (estimated) standard deviation of the observations. The ML estimates $\hat{\boldsymbol{\theta}}$ of the parameters are the values that maximize Equation 3.6, which is equivalent to the minimum of the following function:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{\sigma_{\mathrm{i}}} (\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2 \tag{3.7}$$

as the other terms are constant values. This optimization problem (minimization) is well-known as **weighted least squares**. When the σ_{i} cannot be estimated, they can also be assigned by engineering judgement based on the experimental conditions. This enables the modeller to express the reliability about the measurements in the optimization problem (Dochain and Vanrolleghem, 2001).

In case the standard deviations σ_{i} are assumed constant (homoscedastic), Equation 3.7 further simplifies to the well known SSE performance metric:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2 \tag{3.8}$$

Still, it is important that the usage of SSE within the theoretical framework of least square estimation is only valid if the assumptions of homoscedastic, independent and Gaussian errors are satisfied. The latter is however generally not true for environmental models (Schoups and Vrugt, 2010).

Bayesian Estimation

Whereas the ML estimation assumes the (real) parameters to be constant values and observations are considered as random variables $\boldsymbol{Y}$, in a Bayesian approach the parameters itself are considered as random variables as well. The Bayesian approach updates the prior knowledge (distribution) of the parameters by conditioning it by experimental evidence supporting a continuous learning process.

The parameter prior knowledge is described by the probability $P(\boldsymbol{\theta})$, whereas the information looked for is the knowledge of the parameters conditioned by our available data $P(\boldsymbol{\theta} \mid \mathbf{y})$, called the posterior parameter distribution. Furthermore, the probability (likelihood) $P(\mathbf{y} \mid \boldsymbol{\theta})$ has been defined in the previous section and can be interpreted similarly. The relation in between those terms is provided by the Bayes Theorem:

$$P(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(\mathbf{y})} \tag{3.9}$$

with $P(\mathbf{y})$ the probability density of measured data, which in practice corresponds to a normalization term (VanderPlas, 2014). Hence, Equation 3.9 can be written as:

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \propto P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) \tag{3.10}$$

Direct analytical computation of the posterior parameter distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$ is mostly infeasible. Therefore, the posterior distribution can be approximated numerically by a MCMC sampling scheme, for which different algorithms do exist (section 3.5).

More in-depth knowledge about the theory and application of Bayesian inference is provided by some excellent books on Bayesian statistics, for which MacKay (2002) and Gelman et al. (2013) are of particular interest. For practical application, the

online series of Jupyter notebooks, called *Probabilistic Programming and Bayesian Methods for Hackers* (Davidson-Pilon, 2015) is of great value, since it is completely interactive and reproducible in terms of implementation.

Both frequentistic and Bayesian approaches have their value and the appropriateness of either using a ML (frequentistic) approach or a Bayesian approach is an ongoing debate, which goes beyond the boundaries of environmental science (not to be confused with the discussion of using informal and formal likelihoods in the hydrological community (Beven, 2008a)). However, it is sometimes ignored that under specific conditions, the results are comparable (VanderPlas, 2014). It is important to understand that both are using likelihood functions that represent the assumed model error and can provide complementary information (cfr. section 4.5). For a more extended comparison, the reader is invited to check the online material provided by VanderPlas, which has been published in VanderPlas (2014).

More elaborate likelihood functions intended for environmental modelling are proposed in literature (Kuczera et al., 2006; Schoups and Vrugt, 2010; Renard et al., 2011; Smith et al., 2015). However, the usage of a more elaborate likelihood description goes beyond the (classical) idea of describing the measurement error in the stochastic term. The error model then acts as an additional part of the model structure (Romanowicz et al., 1994). Basically, this extends the entire model description (process model and error model) with a set of extra parameters that need to be inferred as well. Apart from the practical difficulties to estimate an additional set of parameters, adding complexity to the error description could potentially obscure the model structural deficiencies by treating them as stochastic variables in some error term (Gupta et al., 1998). As such, within a diagnostic approach, the focus is given to testability of a wide range of performance metrics (both likelihood functions and other types of metric) rather than the quest for a generally applicable likelihood description.

3.4.3 Including data uncertainty

Besides the limitation present in any model structure, observations are uncertain as well and often do not comply with the assumptions of Gaussian errors. The uncertainty of the observations should be included when this information is available (both in terms of forcing/input data as well as any observations used to perform model evaluation).

A straightforward option, which directly follows from the derivation of the weighted least-squares estimation (section 3.4.2), is using the error information (covariance matrix of the observations) to define the weights of the Weighted Sum of Squared Errors (WSSE) performance metric. In general, lower measurement accuracy of the measurement equipment will be expressed as smaller weights, making them less pronounced in the performance metric. When multiple variables are included, the weights can be used to attribute the reliability of the measurements of the different variables. For example, when the measurement device for a variable is considerably more reliable than the other sensors, higher weights are given. Alternatively, manipulation of the weights of each observation independently is also possible. Ternbach (2005) proposes a function in order to have the standard deviations σ_i of Equation 3.7 (inverse of the weights) proportional to the value of y_i but increasing when the latter approaches the detection limit or the lower accuracy bound of the observed state variable. The pyideas Python Package 2 supports these type of manipulations when defining the measurements (Van Daele et al., 2015c).

Elaborate work is also done in the definition of the error term (likelihood function) within a Bayesian approach. Schoups and Vrugt (2010) propose a likelihood function to cope with correlated, heteroscedastic, and non Gaussian errors taking indirectly the measurement errors into account. This is done at the cost of extra parameters that need to be estimated in the inference process. Smith et al. (2015) define some alternatives with different assumptions leading to different likelihood descriptions. A complementary approach is also the augmentation of the likelihood function by introducing latent variables interacting with the forcing data. When computational resources are available to deal with the extra parameters that need to be inferred, these approaches provide a promising handling of, mostly uncertain, rainfall forcing data (Kavetski et al., 2006a; Renard et al., 2011). Another approach to account for data uncertainty is called *limits of acceptability* (Beven, 2006, 2008b). The proposal came within the scope of the Generalized Likelihood Uncertainty Estimation (GLUE) framework, but it actually can be regarded as an alternative way of constructing performance metrics that can be used in the variety of existing algorithms. Moreover, it is largely similar to the *set-membership* approach as performed by Vanrolleghem and Keesman (1996), using symmetric bounds around the observations without assuming any statistical properties of the errors.

The limits of acceptability approach starts with defining (or assuming) any kind of function that describes the uncertainty of each observation independently. Existing applications used a triangular (Westerberg et al., 2011b; Liu et al., 2009b) or a

trapezoidal (Pappenberger et al., 2006) function. However, it can be a binary function as well, providing 1 within a predefined (uncertainty) region around the observation and 0 outside this region.

Based on the assumed function, the performance of a model simulation is calculated by comparing the individual observations with the corresponding model outputs. For each observation, the equivalent model output is compared and the function defines the performance (i.e. score) of the model simulation for that particular point. In the case of a binary function, it would result in a value 0 or 1 for each of the observations. Combining the scores provides a way to create an aggregated performance metric. The aggregation of these individual scores can be done by summing them up, or alternatively, by expressing it as the number of observations that are approached above a chosen score value.

The approach of limits of acceptability provides a large set of options to construct case-specific performance metrics, taking into account prior knowledge of observational uncertainty. It has been applied mainly in the context of rating curve analysis (Pappenberger et al., 2006; Blazkova and Beven, 2009; Westerberg et al., 2011b), but as well in waste-water treatment modelling (Vanrolleghem and Keesman, 1996). The versatility of this approach makes it appealing within the diagnostic approach.

3.4.4 Combining performance metrics

As mentioned earlier, the application of a single aggregation metric is mostly insufficient to properly characterise model deficiencies (Gupta et al., 1998). However, using multiple performance metrics imposes the question on how to combine the information of the individual performance metrics. The assessment of the individual metrics next to each other is always an option. In function of optimization, assessment of parameter identification and sensitivity analysis, different kinds of combinations are possible as well.

A straightforward approach is to create a single overall metric by combining different metrics into an overall performance metric, e.g. summing them up (or any other aggregation function). When doing so, it is important to keep in mind the magnitude of the individual metrics and whenever possible, to make them relatively comparable. Gupta et al. (2009) proposed the Kling Gupta efficiency (KGE), which computes the Euclidian distance between three important components for model evaluation: correlation, variability error and bias error. Since all of them are dimensionless numbers, the combination by the Euclidian distance

is appropriate. By doing so, the performance metric represents a multi-objective perspective for evaluation.

A similar approach is also executed by Madsen (2000) and van Griensven and Bauwens (2003). By changing the weighting coefficients of individual components, one can obtain alternative solutions. Hence, the applied weights are a subjective choice with direct influence on the end-result and appropriate weighting is therefore essential (Efstratiadis and Koutsoyiannis, 2010). The approach of van Griensven and Bauwens (2003) was later reformulated within a probabilistic approach, which boils down to the multiplication of the individual terms, assuming independent probabilities (van Griensven and Meixner, 2007).

Another approach is to interpret the problem as a multi-objective problem by applying a multi-objective optimization algorithm. This means the characterisation of the pareto front that collects all the optimal combinations of the included performance metrics. Different multi-objective algorithms do exist and are enlisted by Efstratiadis and Koutsoyiannis (2010). Within the metric-oriented approach, existing algorithms such as provided by Fortin et al. (2012), can be coupled by providing the preferred metric functions as input functions to optimize.

When applying a filtering approach, i.e. labelling simulations as behavioural when satisfying a predefined minimal performance (threshold) and considering the others as not behavioural, combining multiple metrics is rather straightforward. New thresholds will diminish the set of behavioural simulations until at some point none of them will satisfy all defined requirements.

A recently proposed technique called Approximate Bayesian Computing (ABC) provides a promising framework to unify the application of multiple performance metrics within a Bayesian approach as a likelihood-free version of parameter inference. Hence, instead of using explicit likelihood functions that are subject to assumptions about the error term, the application of any kind of performance metric could be used to derive information about the posterior parameter distributions. It basically provides the ability to estimate the $P(\boldsymbol{\theta} \mid \mathbf{y})$ of Equation 3.10 based on a set of simulations that applies to a predefined threshold and at the same time provides the ability to use more efficient sampling schemes such as MCMC (Sadegh and Vrugt, 2014). In other words, it provides an efficient approach to the *limits of acceptibility* approach as proposed by Beven (2006) and it is a rigorous framework to handle multiple performance metrics (Vrugt and Sadegh, 2013). This is directly in line with the diagnostic approach presented in this dissertation.

3.5 Sampling strategies

Sampling strategies are a problem of dimensionality and efficiency, needed in environmental modelling due to the impossibility of analytical approaches. However, the necessity of an improved sampling strategy blurs sometimes the communication about the algorithm itself. As considered earlier, when an (almost) infinite availability of model runs would be feasible, a random sampler available in most software environments and (high-level) programming languages would be sufficient in these model applications. However, the dimensionality of the problems typically at hand does require a set of improved sampling strategies, making them indispensable. Hence, current algorithms applied in environmental modelling to perform a sensitivity analysis, uncertainty analysis or optimization typically rely on the application of taking samples from a random variable X described by a PDF f_X in a manner that is mostly directly linked with the methodology.

Sampling strategies are a notable part of statistical research. However, within environmental sciences, practitioners are only aware of a rather limited portion of it. This sometimes leads to a confusion of tongue when talking about the different aspects of random sampling, e.g. mixing up the sampling strategy (how a repeated set of samples is taken) and the used PDFs. In literature and application studies the choice for so-called *uninformative* uniform distributions is commonly seen. This leads to the impression of a common - good - practice of doing so. For some applications the decided distribution sampled is indeed less important than the range of the values sampled (Nossent, 2012). However, when more information is available, the sampling of other distributions, which can be multivariate as well, should be supported too.

Besides, in practical environmental applications, the limitation of conservative random sampling based methodologies are too often undervalued and the necessity of a sufficient amount of samples is regularly ignored. So, next to the ability of sampling all kinds of (multivariate) distributions, the usage of more efficient sampling methods and the verification of convergence are essential (Nossent et al., 2013; Vanrolleghem et al., 2015). A complete overview on the matter is outside the scope of this dissertation, but both issues will be shortly discussed and some practical consequences will be illustrated. This provides the reader with sufficient background to understand the applications in chapter 4, chapter 8 and chapter 10. For an adequate insight on the matter, the reader is referred to Devroye (1986) and MacKay (2002). Both provide a good balance between theoretical background and practical applicability without an environmental specific application.

3.5.1 Sampling non-uniform distributions

Different methods do exist to support sampling from a wide range of distributions, mostly starting with the sampling of a random variable with a PDF from which samples can easily be drawn (e.g. uniform distribution). The *inverse method* uses the inverse of the Cumulative Density Function (CDF) F_X to translate a uniformly sampled value within the interval $[0, 1]$ to a sample of the random variable X, if the inverse F_X^{-1} can be calculated (Figure 3.2). When not, approximated methods do exist as well. As such, this approach can be used for most distributions and implementations of this conversion is common practice in existing statistical packages (e.g. Python Module 1).

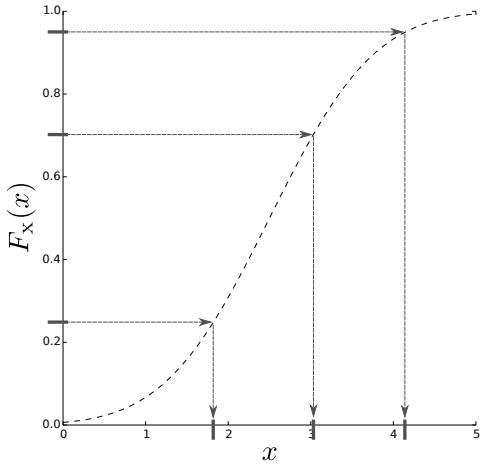


Figure 3.2: Illustration of the sampling of a custom PDF, performed for three samples. By taking the sample in the interval $[0, 1]$ and using the inverse CDF, realisations x of the custom PDF are randomly sampled.

Python Module 1 (scipy.stats)**.**
This module contains a large number of probability distributions, both continuous and discrete. The inverse of the CDF can be calculated by the percent point function (ppf) and as such, a randomly sampled value from a uniform distribution in the interval $[0, 1]$ can be translated in a random sample of the chosen distribution function.

An alternative method is the *acceptance-rejection sampling*, which is less efficient as the *inverse method*, but can be used in the case that the inverse CDF is not known. The *rejection method* is worthwhile mentioning due to its link with the

increased application of MCMC to sample posterior distributions. It samples from a known distribution that encloses the required distribution until a sample is found for which the acceptance criterion is fulfilled. For a general introduction to non-uniform sampling procedures the reader is referred to Devroye (1986).

3.5.2 Sampling strategy

The first step is knowing how to take a single sample from any PDF, based on the sampling from a uniform (the range $[0, 1]$) or other easy to sample distribution. The next step is the strategy to approximate the distribution reliably by sampling the entire range as efficiently as possible (e.g. equal samples of all numbers between 0 and 1). The most straightforward way is the random sampling. However, due to the discrete nature of computers, true randomness cannot be implemented and these methods are referred to as pseudo random sampling such as the Mersenne twister developed by Matsumoto and Nishimura (1998) and used in Python numpy (van der Walt et al., 2011) and Matlab®. Another straightforward approach is to sample following a predefined grid with equal density, which will be perfectly fine for lower dimensional problems, but quickly run into limitations due to the curse of dimensionality when facing higher-order situations (Bergstra and Bengio, 2012).

To overcome the very slow convergence of these pseudo random samplers and the limitations of a grid based approach for more dimensional problems, alternative sampling schemes were developed to improve the coverage. Latin-Hypercube sampling (McKay et al., 1979), where the range is divided in N intervals of equal density 1/N and a (pseudo)-random sample is taken in each interval, is particularly popular due to the ease of implementation. Orthogonal Array-Based Latin-Hypercube sampling (Tang, 1993) provides an improved approach to construct Latin Hypercubes for numerical integration. Quasi-pseudo random sampling approaches actively avoid clustering by taking successive samples away from earlier sampled points, resulting in deterministic sequences that strive for optimal coverage (Nossent, 2012). Of particular interest are Sobol quasi random sampling sequences which are commonly used within the scope of sensitivity analysis (Sobol, 1967; Sobol and Kucherenko, 2005). Moreover, some modelling techniques use a specialized sampling strategy, such as the Morris trajectories within a screening approach for sensitivity analysis (Campolongo et al., 2007) (section 5.3).

The last years, the usage of MCMC sampling is increasingly popular, due to its specific capacity to sample ill-normalized (or otherwise hard to sample) PDFs. MCMC represents a set of methods that are based on sampling values from an approximate

distribution and then correcting those draws iteratively to better approximate the target distribution until convergence is reached (Devroye, 1986). These methods typically use an acceptance-rejection sampler, such as Metropolis-Hastings (Metropolis et al., 1953), to draw new samples from the target distribution (Patil et al., 2010). The Bayesian Total Error Analysis (BATEA) (Kuczera et al., 2006; Kavetski et al., 2006a) and DiffeRential Evolution Adaptive Metropolis (DREAM) (Vrugt et al., 2008a; McMillan and Clark, 2009; Schoups and Vrugt, 2010; Vrugt, 2015) methodologies both rely on an MCMC scheme and are of particular interest in the hydrological community, but both contain one specific implementation of an MCMC approach focusing on high-dimensional problems as it is targeted by other implementations as well (Foreman-Mackey et al., 2013; Patil et al., 2010). Despite the advantages provided by both BATEA and DREAM, the lack of modularity in extracting the sampling scheme itself is a missed opportunity.

Actually, for one or two dimensional problems, a grid based approach would be applicable as well to approximate the posterior distribution. For a detailed overview of the historical development, the theoretical background and the currently existing methodologies for MCMC, the reader is referred to Gelman et al. (2013).

The main advantage, apart from the theoretical considerations, is the efficiency of sampling with a preferential sampling strategy, as the MCMC provides when searching for preferential (optimal) regions in the parameter space. If one would attempt performing a brute force technique by visiting all points in the space and would divide each dimension in 50 equally spaced points, than for a 2-dimensional space it would require $50^2 = 2500$ simulations, but for 10 dimensions this would be $50^{10} = 97656250000000000$ simulations, which is a horrible amount. Hence, the benefit of improved sampling strategies and optimization algorithms should not be underestimated. However, the aim of environmental studies should not be the application of MCMC in itself, but MCMC should be regarded as an efficient tool to sample a distribution.

To summarize, the necessary sampling strategy is dependent on the dimension of the problem, the chosen distribution and the modelling technique itself. In any kind of sampling approach, a proper convergence assessment is essential in order to derive reliable results (Gelman et al., 2013; Nossent et al., 2013; Vanrolleghem et al., 2015). It is noteworthy that approaches for direct multivariate sampling of a known distribution are not explicitly considered here, but should be applied when the information is known. Techniques do exist when the correlation between parameters is known (Iman and Conover, 2007). It is similar to an inverse CDF approach, in the case that the multivariate distribution is described as a copula function, by sampling uniformly in the $[0, 1]$ interval (Vandenberghe, 2012).

3.5.3 Numerical optimization: picking the fast lane

Numerical optimization algorithms are not regularly dealt with at the same time as sampling methods. Mathematically, the purpose is indeed completely different. However, it is considered here as a general way of finding the shortest path from the input (parameter) space to a smaller target space based on the minimization of a defined performance metric, independent of how this metric is constructed. The optimization method to choose depends on the response surface characteristics (resulting from the performance metric selected, the model structure and the data), which can be ranging from a well-known SSE leading to a non-linear least square estimation to a more extended likelihood function leading to maximum likelihood estimation (section 3.4.2) or any metric the modeller can construct to support the model evaluation. However, the optimization itself can be achieved by a variety of existing algorithms and the mathematical properties of the optimization problem are essential to pick a proper optimization algorithm. For a general overview of numerical optimization methods and their respective properties, the reader is referred to Nocedal and Wright (2006).

Similar to sampling methods, optimization problems are of interest to a large area of scientific research. The rather limited set of optimization algorithms applied in environmental research is surprising. Consider for example the popularity of Shuffled Complex Evolution (SCE-UA) as an optimization method regularly encountered in scientific literature of hydrological studies (Duan et al., 1994; Xu et al., 2013; Maier et al., 2015; Willems et al., 2014; Wolfs et al., 2015). The popularity is understandable, since it provides a model structure independent (derivative free), global search algorithm with good convergence properties (Duan et al., 1992). However, the limited number of applications outside this community suggests that many other algorithms could be used as well.

Performance metrics, when implemented as a function, can easily be used as an input argument in existing optimization algorithms such as those provided by the Python Module 2. Within the scope of this dissertation, two specific algorithms are applied to solve different optimization problems, namely gradient based local methods as they are provided by the scipy.optimize.minimize Python Module 2 (Jones et al., 2001) and the implemented SCE-UA Python Module 3.

Python Module 2 (scipy.optimize.minimize)**.**
Minimization of a function of one or more variables, for optimization problems of the form `minimize` $f(\boldsymbol{x})$. Optionally, the lower and upper bounds for each element in $\boldsymbol{x}$ can also be specified using the `bounds` argument. The function provides algorithms for constrained and unconstrained optimization. Both gradient based (Nopens, 2010) as well as gradient free algorithms such as Nelder-Mead (Nelder and Mead, 1965) are available. A good introduction to the different considerations that need to be taken into account (convexity, smoothness and constraints) is given in the scipy lecture notes, continuously updated.
(http://docs.scipy.org/doc/scipy-0.16.1/reference/optimize.html)

Local methods are appropriate when the response surface, i.e. the performance metric in function of the individual elements, is smooth or when working with convex problems. However, the non-linear characteristics of environmental models and the numerical approximations used are causing discontinuities, long ridges and secondary optima, hampering the optimization process (Kavetski and Kuczera, 2007; Schoups et al., 2010). In general, the lack of identifiability of the parameters in a model structure will hamper the success of optimization algorithms for environmental modelling (Andréassian et al., 2012). Multiple combinations of parameter values can provide a similar performance towards a specified metric (Beven, 2002, 2008b).

Python Module 3 (Optimization_SCE)**.**
The SCE-UA algorithm is a global optimization algorithm to find the optimal combination of an input vector $\boldsymbol{x}$ to minimize a function $f(\boldsymbol{x})$ (Duan et al., 1992). It combines the properties of a controlled random search, the Nelder-Mead method (Nelder and Mead, 1965) and competitive evolution (Nossent, 2012). By the simultaneous and independent evolving of different complexes and by regular shuffling in between the complexes, the global minimum is searched for. A detailed description is given in Duan et al. (1992) and Van Hoey (2008).
(https://github.com/stijnvanhoey/Optimization_SCE)

3.6 Conclusion

Many environments and methodologies for model calibration, sensitivity analysis and uncertainty analysis are available in literature. However, the central position of the used metric and the direct link of the metric with the research objective is regularly ignored, leading to the usage of typical metrics over and over again.

By explaining the central role metrics have in any kind of model evaluation algorithm, the chapter provides a common denominator for many of the existing methodologies. Making a clear division in between the analysis itself (e.g. identifiability, uncertainty, sensitivity analysis) and the necessity of a sampling technique is crucial, but regularly ignored.

The chosen metric is the link between the algorithm and the research question and should be of major concern in the first place. Based on the metric definition, the necessity of a formal framework or any kind of algorithms for numerical approximation follows, not the other way around. By making this clear to future modellers and practitioners, the chapter aims to support the objective of improving current practices in model evaluation.

CHAPTER 4

Case study: respirometric model with time-lag

Parts redrafted and compiled from

Cierkens, K., Van Hoey, S., De Baets, B., Seuntjens, P., and Nopens, I. (2012). Influence of uncertainty analysis methods and subjective choices on prediction uncertainty for a respirometric case. In Seppelt, R., Voinov, A. A., Lange, S., and Bankamp, D., editors, *International Environmental Modelling and Software Society (iEMSs) 2012 International Congress on Environmental Modelling and Software. Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting*, Leipzig, Germany. International Environmental Modelling and Software Society (iEMSs)

Decubber, S. (2014). *Linking the carbon biokinetics of activated sludge to the operational waste water treatment conditions.* Msc thesis, Ghent University

4.1 Introduction

Different diagnostic tools can be used to evaluate model structures and collect information to support the model calibration. The previous chapter provided a more general background and introduced the concept of the metric oriented approach, putting the choice of metrics first. In this chapter, an ODE-based model focusing on respirometry will be used to provide a practical application of the suggestions made in the preceding chapter.

The aim of the modelling exercise at hand is defined as 'to evaluate the identifiability and applicability of a chosen respirometric model structure for aerobic

degradation of acetate based on ASM No. 1'(Henze et al., 1983; Gernaey et al., 2002). Vanrolleghem et al. (1995) and Dochain et al. (1995) studied the identifiability of a similar model structure, without the addition of a time-lag function representing the retardation of the biomass activity. The analysis in this chapter aims to check the correspondence with the earlier work and to evaluate if the addition of the transient term can be justified as a model structure decision.

Three different approaches will be used to perform the analysis:

- The influence of the experimental conditions on the parameter identifiability is questioned. As previous research stated that the initial amount of acetate should be appropriate to satisfy the assumptions of the model structure (Grady et al., 1996), the effect of the ratio substrate to biomass will be evaluated. As we are specifically interested in the influence of the time-lag function, the parameter influence as a function of time is assessed. While choosing each individual time step as a metric, a **local parameter sensitivity analysis** is applied to check the relative sensitivity and interaction of the parameters under the two experimental conditions.
- To have a more general understanding of the influence of the parameters relative to the initial acetate concentration, a **Sobol global parameter sensitivity analysis** is applied. Instead of focusing on a particular point in the parameter space, the influence and interactions are checked for the entire range of the different factors (parameters and initial concentration). Different aggregation metrics are used to compare the influence of the input factors and check the identifiability.
- A normal distribution is assumed for the residuals, leading to a performance metric defined by the SSE. These assumptions enable to use a metric as an estimator in a formal framework (cfr. section 3.4.2). This performance metric is used to **calibrate** the respirometric model by using an ML approach and by sampling the posterior of the proposed likelihood function with an MCMC sampler. The parameter interactions are evaluated under the given assumptions.

To not overload the dissertation itself with redundant code, the execution was performed in a set of Jupyter notebooks [1]. These notebooks provide an interactive environment to reproduce the implementations executed in this chapter.

The chapter is organised as follows. First, the concept of respirometry is briefly introduced for the unfamiliar reader, along with the available observations and the chosen model structure. Next, the three aforementioned analyses are be performed

[1] https://github.com/stijnvanhoey/phd_ipynb_respiro_showcases

and explained. Finally, the conclusions about the suitability of the model structure are collected based on the combined results. At the same time, the chapter illustrates how model evaluation can be based on the local sensitivity as a function of time, based on aggregated metrics or based on metrics enabling a formal approach.

4.2 Respirometry

Respirometric experiments are typically used to characterise aerobic degradation by the active microorganisms in activated sludge (Gernaey et al., 2002). During a respirometric experiment, an amount of biodegradable substrate, e.g. acetate, is added to a batch reactor containing activated sludge. By monitoring the amount of oxygen per unit of volume and time that is consumed by the microorganisms, the respiration rate of the activated sludge can be assessed. It can be applied for carbon source degradation processes, but also for nitrification, i.e. the oxidation of ammonium to nitrate, as well. In model studies of waste water treatment plants, respirometry is applied to estimate biokinetic parameters describing the activated sludge characteristics. As such, the number of parameters to calibrate in the (over-parameterized) ASMs used in full-scale modelling studies can be reduced (Spanjers and Vanrolleghem, 1995; Vanrolleghem et al., 1999). Besides, respirometry is also applied to quantify the different Biological Oxygen Demand (BOD) fractions in waste water and to evaluate toxicity of waste water. The focus is on the characterisation of biokinetic parameters by evaluating simplified ASM models based on the observed experimental data.

The parameter identification of models used within the scope of respirometry has been studied and described extensively in the past (Dochain et al., 1995; Vanrolleghem et al., 1995; Spanjers and Vanrolleghem, 1995; Grady et al., 1996; Vanrolleghem et al., 1999; Petersen et al., 2001; Gernaey et al., 2002; De Pauw, 2005). Most of the strategies to overcome the lack of identifiability are in line with the idea of extracting additional information out of observations to support parameter identification, both by measuring specific parameters individually (Vanrolleghem et al., 1999) or by extracting information from additional data sources, e.g. titrimetric data (Gernaey et al., 2002). Sensitivity analysis is in many cases an essential element in the assessment of the parameter identifiability.

It is important to understand that on the basis of the oxygen uptake rate or the measured dissolved oxygen levels alone only a subset of parameters are structurally identifiable (Dochain et al., 1995). Moreover, practical identifiability issues

are reported as well. For example, Vanrolleghem et al. (1995) describe the interaction between the maximum growth rate μ_{max} and the half saturation coëfficient K_S when other parameters and initial conditions are assumed known leading to ill-conditioned parameter estimates. Figure 4.1, taken from Vanrolleghem et al. (1995), illustrates this specific interaction effect. As such, the respirometry application provides a well known case study to showcase different model evaluation strategies and to explore the parameter sensitivity/identifiability tools.

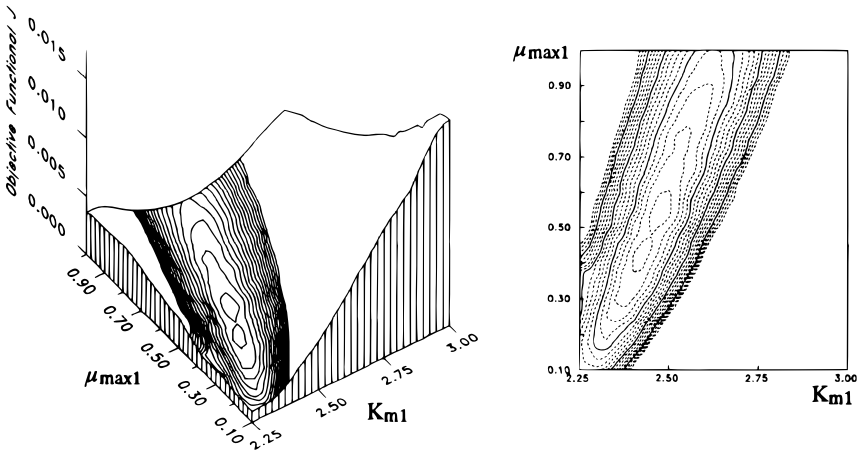


Figure 4.1: Resulting figure of a practical identifiability analysis performed by Vanrolleghem et al. (1995), showing parameter interaction between parameters μ_{max} and K_S, the latter named K_{m1} in the original paper. (figure reproduced from Vanrolleghem et al. (1995))

4.2.1 Respirometric data collection

A first set of experiments used in this work is described in Cierkens et al. (2012). The flowing gas-static liquid respirometer consists of a reactor with a volume of 2 l filled with sludge, taken from the aerobic tanks of the municipal WWTP of Ossemeersen (Gent, Belgium). The sludge was aerated overnight to ensure endogenous state. Temperature is controlled at 20 °C (± 0.05) and pH at 7.5 (± 0.1). Dissolved oxygen and pH are recorded every second with an LDO sensor (Mettler Toledo, Inpro 6870i) and a pH-sensor (Mettler Toledo HA 405-DXK-S8/225). An acetate pulse of 60 $mg\,l^{-1}$ was added according to Gernaey et al. (2002). Exogenous oxygen uptake rate (OUR_{ex}) profiles are calculated similar to Petersen (2000). Figure 4.2 visualizes the observed dissolved oxygen concentration S_O and the calculated OUR_{ex} of a single experiment. Hence OUR_{ex} is a derived metric instead of a direct observed value.

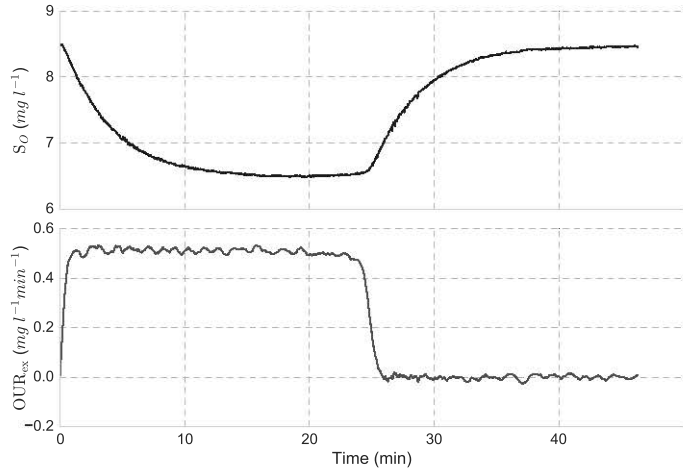


Figure 4.2: Observations from a single respirometric experiment as performed by Cierkens et al. (2012). The top graph represents the measured oxygen concentration S_O ($mg\,l^{-1}$) and the lower graph represents the exogenous oxygen uptake rate: OUR_{ex} ($mg\,l^{-1}\,min^{-1}$)

Additional experiments were performed by Decubber (2014) with a similar experimental setup with sludge taken from a full-scale A/B-installation of Nieuwveer (Breda, Netherlands). Instead of a single acetate pulse of 60 $mg\,l^{-1}$, individual experiments consisted of dosing consecutive substrate spikes, each time when the OUR had dropped back to endogenous levels with changing levels of acetate dosage. A detailed description of the experiments and the experimental setup is provided in Decubber (2014). Figure 4.3 provides the outcome of a single experiment (reference number 0508A), showing the acetate spikes and the resulting drops in dissolved oxygen S_O caused by the microorganisms activity.

4.2.2 Respirometric model

A simple respirometric model for aerobic degradation (Equations 4.1 till 4.4) of acetate S_A without storage is used (Gernaey et al., 2002), based on ASM No. 1 (Henze et al., 1983). It predicts the exogenous oxygen uptake rate OUR_{ex} ($mg\,l^{-1}\,min^{-1}$), caused by the substrate (in this case acetate) consumption by the active biomass X_B to grow following Monod kinetics. Endogenous respiration of the biomass, i.e. the basal metabolism of the biomass is also described. The OUR_{ex} is derived using an algebraic equation, based on the dissolved oxygen state variable S_O. A

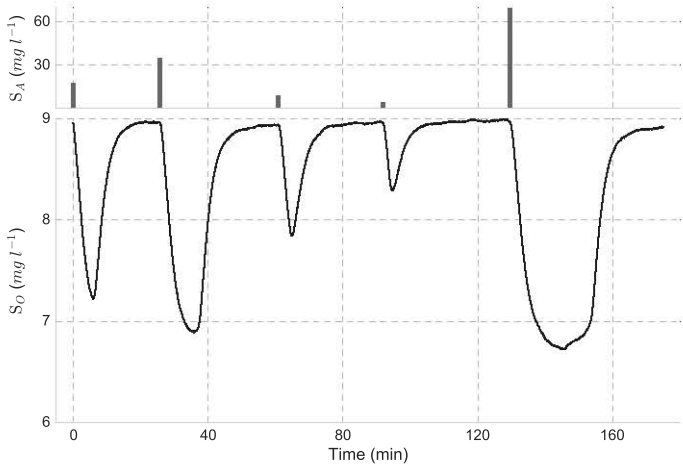


Figure 4.3: Observations from a single respirometric experiment (ref. id 0508A) as performed by Decubber (2014). The top graph represents the added acetate S_A ($mg\,l^{-1}$) and the lower graph represents the measured oxygen concentration S_O ($mg\,l^{-1}$). This experiment consisted of five consecutive acetate spikes.

model representation is also provided in Table 4.1 as a Gujer matrix, providing a standardised model representation, common for ASM model descriptions. The matrix representation is equivalent to Equations 4.1 till 4.4. A more detailed explanation of this matrix model representation is provided in section 9.4. Table 4.2 provides an overview of the different state variables, parameters and initial conditions.

$$\frac{dS_A}{dt} = -(1 - e^{-\frac{t}{\tau}})\frac{1}{Y}\mu_{max}\frac{S_A}{K_S + S_A}X_B \tag{4.1}$$

$$\frac{dX_B}{dt} = (1 - e^{-\frac{t}{\tau}})\frac{1}{Y}\mu_{max}\frac{S_A}{K_S + S_A}X_B - b\,X_B \tag{4.2}$$

$$\frac{dS_O}{dt} = -(1 - e^{-\frac{t}{\tau}})\frac{1-Y}{Y}\mu_{max}\frac{S_A}{K_S + S_A}X_B - b\,X_B + k_{La}(S_O^* - S_O) \tag{4.3}$$

$$OUR_{ex} = (1 - e^{-\frac{t}{\tau}})\mu_{max}\frac{1-Y}{Y}\frac{S_A}{K_S + S_A}X \tag{4.4}$$

A typical observation in short-term batch experiments such as a respirometer, is that the respiration signal exhibits a transient response before attaining its maximum value (Vanrolleghem et al., 2004). This time lag is typically of the order

Table 4.1: Representation of the respirometry model as a Gujer matrix consisting of state variables to represent aerobic degradation of acetate S_A by biomass X_B consuming oxygen S_O. Units of the state variables are expressed as $M_{(COD)}l^{-1}$ to explain the stochiometric correspondance, which is equivalent to $mg\,l^{-1}$ using the molar mass of each substance.

process	stoichiometry			reaction rate
	X_B	S_A	S_O	
heterotrophic growth with S_A as substrate	1	$-\frac{1}{Y}$	$-\frac{1-Y}{Y}$	$(1-e^{-\frac{t}{\tau}})\mu_{max}\frac{S_A}{K_S+S_A}X_B$
endogenous respiration	-1		-1	$b\ X_B$
aeration			1	$k_{La}(S_O^*-S_O)$
stochiometric parameters: Y	biomass $(M_{(COD)}l^{-1})$	substrate $(M_{(COD)}l^{-1})$	oxygen $(M_{(-COD)}l^{-1})$	kinetic parameters: μ_{max}, K_S, τ k_{La}, S_O^*, b

of minutes and therefore only observed when the frequency of S_O measurement is of the order of seconds. Vanrolleghem et al. (2004) suggested that the time lag, which is of the order of minutes, can only partially be accounted for by the dynamics of the oxygen sensor and by improper mixing. Neither can it be explained by diffusion limitation of oxygen into the sludge flocs, since similar time lags have been observed in experiments with dispersed single species cultures. They suggested that the time lag can be further explained by intracellular phenomena such as delays in substrate metabolism and concluded that it can be described by a first-order model of the growth rate with following term: $(1-e^{-\frac{t}{\tau}})$. τ is the time coefficient that needs to be determined in order to correctly describe the retardation of the biomass activity.

Dochain et al. (1995) studied the structural identifiability of the considered model focusing on the Monod kinetics, i.e. without the transient time lag function, oxygen transfer and ignoring the biomass decay. Hence, from the five remaining parameters μ_{max}, K_S, Y and S_A^0 (they considered the initial concentration as a parameter for the analysis), only the following three combinations were identifiable: $\mu_{max}X_B(1-Y)/Y$, $(1-Y)S_A^0$ and $(1-Y)K_S$. By assuming X_B, Y and S_A^0 known a

priori, estimation of only μ_{max} and K_S remained in their subsequent paper focusing on practical identifiability (Vanrolleghem et al., 1995).

Table 4.2: Overview of the parameters and states in the used respirometric model

Variable	Description	Units
S_A	acetate concentration	$mg\,l^{-1}$
S_O	dissolved oxygen concentration	$mg\,l^{-1}$
X_B	biomass concentration	$mg\,l^{-1}$
OUR_{ex}	oxygen uptake rate	$mg\,l^{-1}\,min^{-1}$
Parameter		
μ_{max}	maximum growth rate	d^{-1}
K_S	half-saturation Monod constant	$mg\,l^{-1}$
τ	retardation of biomass activity time coëfficient	d
Y	yield of the biomass	-
k_{La}	volumetric gas/liquid mass transfer coefficient for oxygen	min^{-1}
b	biomass decay rate	d^{-1}
S_O^*	saturated oxygen concentration[a]	$mg\,l^{-1}$
Initial condition		
S_O^0	initial oxygen concentration[a]	$mg\,l^{-1}$
S_A^0	initial acetate concentration	$mg\,l^{-1}$
X_B^0	initial biomass concentration	$mg\,l^{-1}$

[a] since the reactor is saturated with oxygen at the start of each experiment, S_O^* is assumed to be equivalent to S_O^0

As suggested by Grady et al. (1996), the experimental conditions within the work of Decubber (2014) are according to S_A^0/X_B^0 ratios that are not altering the community structure (below 0.025). In the oxygen mass balance, the theoretical saturation concentration S_O^* was replaced by the measured equilibrium concentration in order to express the mass balance in terms of the exogenous oxygen uptake rate OUR_{ex} (Decubber, 2014). By saturating the reactor with oxygen till equilibrium is reached prior to the acetate addition, the measured equilibrium concentration is equivalent to the initial concentration S_O^0 of the simulation. In other words, the measured oxygen concentration at equilibrium S_O^0 is used as saturated oxygen concentration S_O^* and assumed known.

Initial concentrations of acetate S_A^0 are controlled as an experimental condition. The initial biomass concentration X_B^0 derivation depends on the model and experimental application. Structural identifiability analysis clarified the impossibility of estimating μ_{max} and X_B^0 separately (Dochain et al., 1995). Cierkens et al. (2012) also illustrated that defining one or the other is needed. As such, to define the biomass, the volatile suspended solids were measured and the assumption is made that X_B^0 is half of this concentration (Decubber, 2014).

Furthermore, the parameters Y, k_{La} and b were experimentally defined or derived from earlier optimizations and the assumed values are enlisted in Table 4.3 (Decubber, 2014; Cierkens et al., 2012). Hence, focus is on the estimation of μ_{max} and K_S in correspondence to Vanrolleghem et al. (1995), but extended with the parameter τ.

Table 4.3: Overview of the parameter values described in respectively Cierkens et al. (2012) and (Decubber, 2014) and here assumed as known values

Parameter	Cierkens et al. (2012)	Decubber (2014)	Unit
Y	0.78	0.70	-
k_{La}	0.26	0.42	min^{-1}
b	0.62	0.24	d^{-1}
S_O^*	8.40	8.94	$mg\,l^{-1}$

The respirometric model was implemented in the Python programming language by using the Python Package pyIDEAS 2 developed to support the creation of any kind of ODE based model represented by Equation 2.1.

Python Package 2 (pyideas)**.**
This pyideas model environment is an object oriented python implementation for model building and analysis, focussing on identifiability analysis and optimal experimental design. It provides a simplified syntax to define general models for ODE models used in this dissertation.

(https://github.ugent.be/pages/biomath/biointense/

Figure 4.4 provides the outcome of a respirometric model simulation. The acetate dosed is consumed by the biomass according to the Monod kinetics, causing the dissolved oxygen concentration to drop and biomass to grow. When the acetate is consumed, the biomass concentration decreases due to biomass decay according

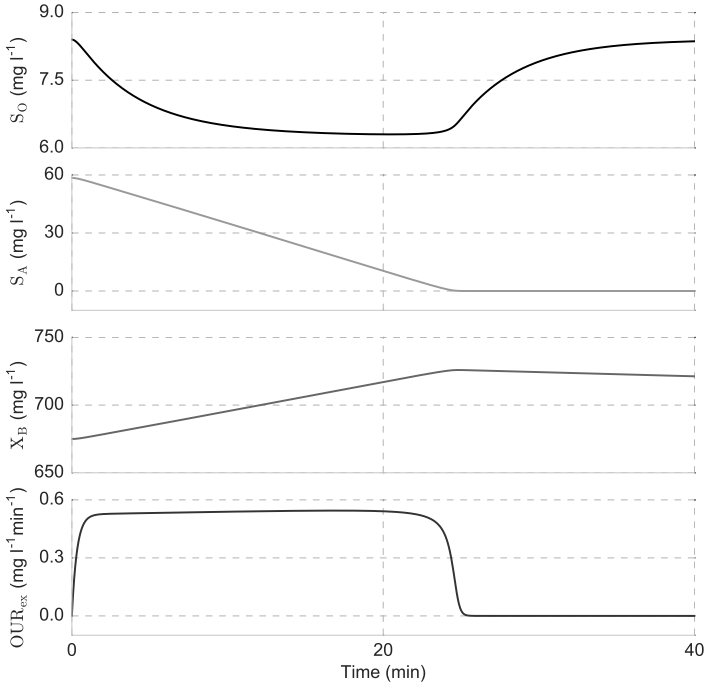


Figure 4.4: Output from a single respirometric model simulation, using $\mu_{\max} = 4\ \mathrm{d}^{-1}$, $K_S = 0.4\ \mathrm{mg\,l^{-1}}$, $\tau = 2.25 \times 10^{-4}\ \mathrm{d}$, $Y = 0.78$, $b = 0.62\ \mathrm{d}^{-1}$, $k_{La} = 0.26\ \mathrm{min}^{-1}$, $S_O^0 = S_O^* = 8.4\ \mathrm{mg\,l^{-1}}$, $S_A^0 = 58.48\ \mathrm{mg\,l^{-1}}$ and $X_B^0 = 675\ \mathrm{mg\,l^{-1}}$. The model calculates the dissolved oxygen S_O ($\mathrm{mg\,l^{-1}}$), the concentration of acetate S_A ($\mathrm{mg\,l^{-1}}$), the biomass concentration X_B ($\mathrm{mg\,l^{-1}}$) and the exogenous oxygen uptake rate OUR_{ex} ($\mathrm{mg\,l^{-1}\,min^{-1}}$). Acetate is consumed by the biomass, whereafter the biomass returns to endogenous activity and the oxygen level increases again.

to the biomass decay rate b and oxygen levels increase again. In the next sections, the model will be used as an example case to illustrate the application of both a local and a global sensitivity analysis.

4.3 Comparing experimental conditions

Introduction to local sensitivity analysis

Sensitivity analysis provides an estimate on the influence input factors have on the model output. A local sensitivity analysis is based on the partial derivatives from a response variable of interest $\hat{y}_i$ to an input factor θ_j:

$$\frac{\partial \hat{y}_i(\boldsymbol{\theta}, t)}{\partial \theta_j} \tag{4.5}$$

taking into account the notation of Equation 2.1. When direct operation on the differential equations are possible, the local sensitivity analysis can be derived analytically, as introduced by Vanrolleghem et al. (1995); Dochain and Vanrolleghem (2001); Donckels (2009). More recently Van Daele et al. (2015c) implemented this based on the symbolic computation provided by SymPy Development Team (2014).

The analytical derivation can become very complicated for complex models and many environmental models do not allow direct operations on the model equations. In these cases, Equation 4.5 has to be approximated using numerical techniques. An overview of the existing methods is given by De Pauw and Vanrolleghem (2006) and a more in depth evaluation is provided by De Pauw (2005). The most straightforward approach is the finite difference approximation, where the sensitivity of the variable $\hat{y}_i$ to parameter θ_j is approximated as

$$\frac{\partial \hat{y}_i(\boldsymbol{\theta}, t)}{\partial \theta_j} = \lim_{\Delta\theta_j \to 0} \frac{\hat{y}_i(\boldsymbol{\theta} + \Delta\theta_j, t) - \hat{y}_i(\boldsymbol{\theta}, t)}{\Delta\theta_j} \tag{4.6}$$

with $\hat{y}_i(\boldsymbol{\theta} + \Delta\theta_j, t)$ the value of $\hat{y}_i$ at time step t when $\Delta\theta_j$ is added to the value of parameter θ_j. A sufficiently small value for $\Delta\theta_j$ should be used to make sure Equation 4.6 is valid. When $\Delta\theta_j$ is chosen too large, the non-linearities of the model will influence the parameter sensitivity calculation and the finite difference approximation will not be valid. However, the numerical accuracy of the model solver restricts the accuracy of the approximation by defining a lower limit. Moreover, when environmental models are used that communicate with input/output text-files, the used floating point number representation needs to be taken into account. Hence, a balance in between the numerical approximation and the practical feasibility has to be found (De Pauw, 2005).

Equation 4.5 provides the calculation of the *absolute sensitivity* since it is dependent on the absolute values of both the parameter and the variable. This limits the possibility of comparing different sensitivity functions with one another. This can be tackled by calculating the *relative sensitivity* towards the variable, parameter or both, called relative sensitivity, parameter relative sensitivity and total relative sensitivity respectively:

- relative sensitivity (RS)
$$\frac{\partial \hat{y}_i(\boldsymbol{\theta}, t)}{\partial \theta_j} \cdot \frac{1}{\hat{y}_i(\boldsymbol{\theta}, t)}$$
- parameter relative sensitivity (PRS)
$$\frac{\partial \hat{y}_i(\boldsymbol{\theta}, t)}{\partial \theta_j} \cdot \theta_j$$
- total relative sensitivity (TRS)
$$\frac{\partial \hat{y}_i(\boldsymbol{\theta}, t)}{\partial \theta_j} \cdot \frac{\theta_j}{\hat{y}_i(\boldsymbol{\theta}, t)}$$

By plotting the sensitivities as a function of time, periods of high sensitivity can be identified. Hence, these can be used to improve the confidence of the parameter estimation. By this property, the local sensitivity is the central element in the proposal of new experiments within an optimal experimental design context (Donckels, 2009; De Pauw, 2005). Comparison of multiple parameter sensitivities in time provides insight into potential interactions when similar individual effects (same or inverse patterns) are observed. As such, the parameter identification is supported during periods of high sensitivity when changes due to a change of a single parameter is not cancelled out by the others. The latter can be quantified by a collinearity analysis providing information about the linear dependencies between model parameters for a specific point in the parameter space (Brun et al., 2001; De Pauw et al., 2008).

Application of the local sensitivity analysis

The aim of the local Sensitivity Analysis (SA) is to examine parameter identifiability and the impact of changes in the parameters μ_{max}, K_S and τ on the model output. To investigate how the application of different substrate additions influences the model behaviour, focus is given on two respirograms with a substrate to biomass ratio S_A^0/X_B^0 of 1/100 and 1/40, respectively (second last and last addition on Figure 4.3). These two respirograms will be further referred to as the

low (1/100) and high (1/40) S_A^0/X_B^0 cases. The implementation of Van Daele et al. (2015c) has been used to perform the local sensitivity analysis.

The point in parameter space used for the sensitivity analysis is determined by a preliminary parameter estimation performed on the individual respirograms (Decubber, 2014). For the high S_A^0/X_B^0 ratio, the resulting parameter values were $\mu_{\max} = 3.78\ \mathrm{d}^{-1}$, $K_S = 1.6\ \mathrm{mg\,l}^{-1}$, $\tau = 9 \times 10^{-4}\,\mathrm{d}$. For the low S_A^0/X_B^0 ratio, the resulting optimal parameter values were $\mu_{\max} = 7.44\ \mathrm{d}^{-1}$, $K_S = 3.17\ \mathrm{mg\,l}^{-1}$, $\tau = 18 \times 10^{-4}\,\mathrm{d}$. The other values were assumed fixed and identical in both cases and enlisted in Table 4.3. The completely different optimal parameter values of both respirograms illustrate the identifiability problem. Since both respirograms are coming from the same experiment and are carried out with the same activated sludge, the kinetic characteristics (parameters) should be the same.

The central total relative sensitivity (TRS) functions for both were evaluated (section 4.3), to enable comparison amongst different parameters (and variables). A perturbation factor of 10^{-4} was used. Figure 4.5 shows the model output of the dissolved oxygen S_O together with the obtained sensitivity functions for the parameters when a high S_A^0/X_B^0 ratio is applied experimentally. Figure 4.6 shows a comparable output when a low S_A^0/X_B^0 ratio is used.

During the declining and constant S_O phase, the sensitivity function for μ_{max} is negative. Indeed, a higher growth rate will correspond to a higher maximum OUR_{ex} which in turn will cause the S_O curve to reach lower oxygen concentrations (i.e., lower model output). The opposite is true for K_S: according to the Monod kinetics, a higher value for K_S means a lower OUR_{ex} at the same substrate concentration.

In the rising S_O region, the sensitivity functions for μ_{max} and K_S switch signs. In this phase, the sensitivity towards an increase in μ_{max} is positive. A higher maximum growth rate for the same substrate concentration means that the substrate will be depleted faster. Therefore, the S_O will rise earlier in time compared to an experiment with lower μ_{max}, or in other words, during the rising S_O phase the oxygen concentration will be already higher at the same time instant for a higher μ_{max}. Following the same logic, the sensitivity function for K_S is explained, except for the fact that an increase in K_S has the exact opposite effect on model output as an increase in μ_{max}.

The model shows a positive sensitivity for the time lag τ during the declining and constant phase. For larger values of τ, the OUR_{ex} response will be slower. This means that for larger values of τ, the OUR_{ex} will be lower at the same time instant (rises more slowly) resulting in a higher S_O concentration, which explains

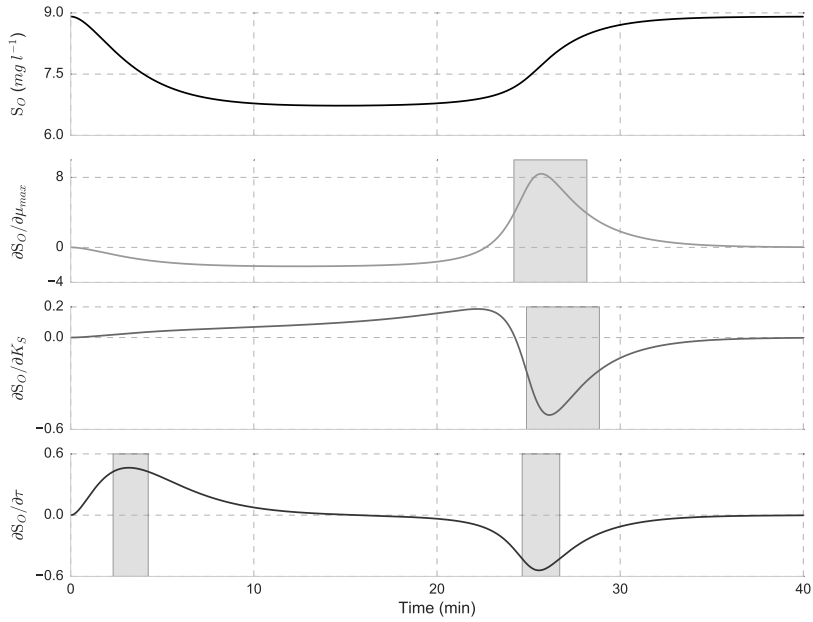


Figure 4.5: Total relative sensitivity functions of the parameters μ_{max}, K_{S} and τ for the dissolved oxygen concentration $\mathrm{S_O}$ with a **high** $\mathrm{S_A^0/X_B^0}$ ratio. The top figure represents the modelled output of $\mathrm{S_O}$ and the other plots respectively the relative sensitivity functions of μ_{max}, K_{S} and τ. The grey highlighted regions are for each parameter the periods with the highest sensitivities, i.e. sensitivities exceeding the 90 % interval of the absolute values. The first highlighted section of $\partial S_O/\partial \tau$ is distinct from the other parameter uncertainties, supporting the identification of the τ parameter value.

the positive sensitivity. During the rising $\mathrm{S_O}$ phase the sensitivity function for τ becomes negative. Indeed, for higher values of τ, $\mathrm{S_O}$ will start rising later in time.

In both cases, the sensitivity of μ_{max} is dominant during the simulation. The sequence of the declining, constant and the rising phases occur in both respirograms, and the explanation for the behaviour of the sensitivity functions is the same for both cases. However, the timing and duration of these characteristic phases is different between the two respirograms. Because of this, in the high $\mathrm{S_A^0/X_B^0}$ case the declining and the rising $\mathrm{S_O}$ phase are separated in time whereas in the low case the second phase immediately follows the first one.

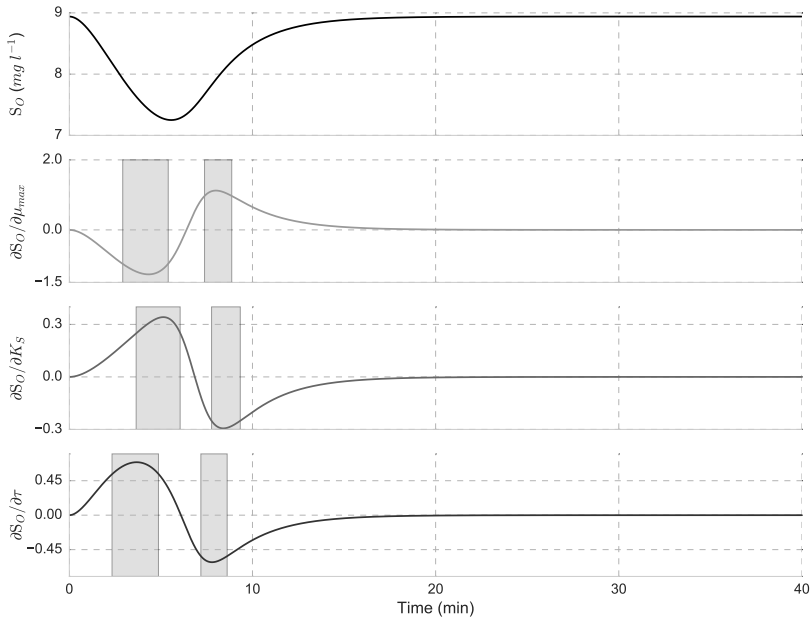


Figure 4.6: Total relative sensitivity functions of the parameters μ_{max}, K_{S} and τ for the dissolved oxygen concentration $\mathrm{S_O}$ with a **low** $\mathrm{S_A^0/X_B^0}$ ratio. The top figure represents the modelled output of $\mathrm{S_O}$ and the other plots respectively the relative sensitivity functions of μ_{max}, K_{S} and τ. The grey highlighted regions are for each parameter the periods with the highest sensitivities, i.e. sensitivities exceeding the 90 % interval of the absolute values. Highlighted sections are during similar time periods for all three parameters, suggesting identifiability issues.

These differences in respirometric curve shape resulting from alternative experimental conditions have some important implications for the sensitivity functions (and related identifiability (Grady et al., 1996)). Both in Figure 4.5 and Figure 4.6, the periods of the largest values (in absolute value) of the sensitivities are marked with grey. In the low $\mathrm{S_A^0/X_B^0}$ case the sensitivity peaks of the individual parameters do overlap, leading to interaction and compensation. In other words, decreasing the identifiability of the parameters during these periods. However, in the high $\mathrm{S_A^0/X_B^0}$ case (Figure 4.5) the first peak sensitivity for τ is separated in time from the peak sensitivities of the other parameters. Notwithstanding that the sensitivity for μ_{max} is still reaching to much larger values, the separation supports the identification of parameter τ.

Hence, the experimental condition does support the identification (and the confidence of estimation) of the model parameters for the selected model structure. At the same time, the analysis provides information about the proposal of new (optimal) experiments as the high S_A^0/X_B^0 case is preferred. The latter is the central theme of OED. The proposal of new experiments is not discussed in this dissertation, considering the available set of observations as the information to work with. However, research focusing on OED has been done by Donckels (2009), De Pauw (2005) and is ongoing (Van Daele et al., 2015b,c).

4.4 Global sensitivity analysis

When interest is on the entire parameter space in combination with other input factors, the application of a global SA is preferred. Since the simulation time to run the respirometry model is short and a quantitative statement about the sensitivity is aimed for, the application of a Sobol sensitivity analysis is chosen.

For readers who are not familiar with the Sobol sensitivity analysis, an in depth explanation is provided in section 5.6. However, for the moment it is important to understand that the Sobol method provides an estimate of the influence of the input factors on the defined model output metric, based on a large set of sampled parameter combinations. Two sensitivity metrics are provided by the analysis: the first order sensitivity index and the total sensitivity index for each of the input factors. The former provides an estimate of the influence of the individual factor on the output, whereas the latter provides an estimate of the influence of each factor together with the interactions this factor has with other factors.

As seen during the local sensitivity analysis, the experimentally chosen substrate addition is important on the resulting output and will have an effect on the identification of the parameters. Accounting for it in the sensitivity analysis provides the possibility of assessing the relative importance of the chosen substrate addition compared to the sensitivity of the individual parameters.

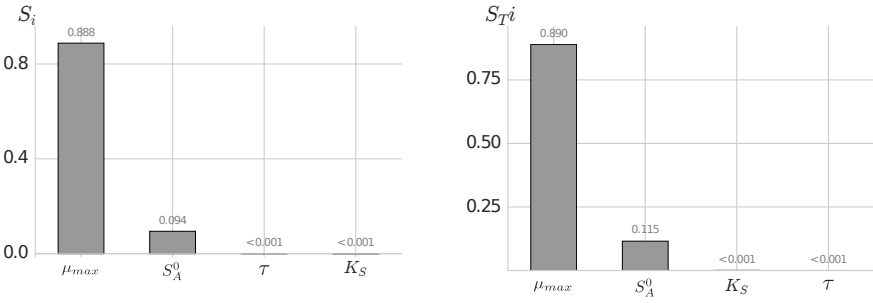
Hence, the value of S_A^0 will be included in the performance of the Sobol sensitivity analysis, ending up with a total of four input factors: μ_{max}, K_S, τ and S_A^0. The experimental conditions and the assumed parameter distributions (i.e. uniform ranges) are taken from Cierkens et al. (2012) except for the τ value. The range of τ was expanded in order to agree with the range of τ values enlisted in Vanrolleghem et al. (2004).

By taking $N = 3000$ base runs will result in a total of $N(k+2) = 3000 \cdot (4+2) = 18000$ model simulations that need to be performed.

To evaluate the sensitivity of the model output regarding the four input factors, a decision needs to be made about the used aggregation metric. The variable to check the sensitivity for will be the oxygen concentration S_O. Still, different options do exist to aggregate the variable.

Importance of chosen aggregation metric

To illustrate the importance of the chosen aggregation metric, Figure 4.7 shows the difference between the influence of the input factors when considering two straightforward options: (1) the reached minimum of S_O (Figure 4.7a) and (2) the average of S_O (Figure 4.8). The first and total indices when using the reached minimum as metric are provided in respectively Figure 4.7a and Figure 4.7b.

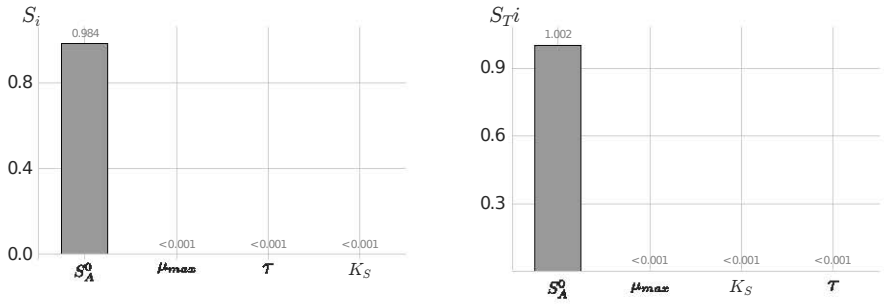


(a) First-order sensitivity index **(b)** Total-order sensitivity index

Figure 4.7: Overview of the sensitivity when choosing the minimum values reached during the simulation as the aggregated metric, considering the four input factors μ_{max}, K_S, τ and S_A^0, for which μ_{max} is most sensitive, mainly by its direct effect on the minimum value.

The calculation of the first and total sensitivity indices using the average value of each simulation are provided in respectively Figure 4.8a and Figure 4.8b.

The importance of μ_{max} when focusing on the minimum value is understandable since it defines the maximum rate at which the active biomass is degrading the substrate, influencing the related oxygen levels (and the uptake rate). Since the amount of substrate added influences the length and the entire shape of the oxygen concentration profile, the sensitivity is largest when focusing on the average value.



(a) First-order sensitivity index **(b)** Total-order sensitivity index

Figure 4.8: Overview of the sensitivity when choosing the average values of the simulations as the aggregated metric, considering the four input factors μ_{max}, K_S, τ and S_A^0, for which S_A^0 is most sensitive, due to the overall effect on the entire S_O profile. The values of S_i and S_{T_i} should be both regarded as 1. The difference is caused by the numerical approximation of the sensitivity indices.

Notwithstanding the rather trivial results, it is clear that the chosen aggregation metric has a direct effect on the result. Hence, there is not such a thing as *the sensitivity analysis of model A*, since it embraces a huge variety of options, each for specific purposes. Hence, a SA should not be the purpose in itself, but rather a tool to answer research questions or to support model structure understanding and evaluation. The latter is regularly ignored in literature.

Check for convergence of indices

It is generally known that variance-based methods for sensitivity analysis require a large set of simulations to let the metrics converge to a proper estimate. Extreme values in the calculated metric values will complicate the convergence of the sensitivity indices (Nossent and Bauwens, 2012a,b). Therefore, checking the convergence is essential for any type of sensitivity method. This can be done graphically by plotting the estimated indices for the first and total variance progressively in function of the performed set of base runs, as provided in Figure 4.9 for the mean value first order sensitivity indices S_i.

After each reconsideration of the used aggregation (or performance) metric the convergence for both the first and total order sensitivity indices can be recalculated and visualised directly. In this specific case, the user could decide to put the number of base runs to 1500 and still provide quantitatively meaningful results.

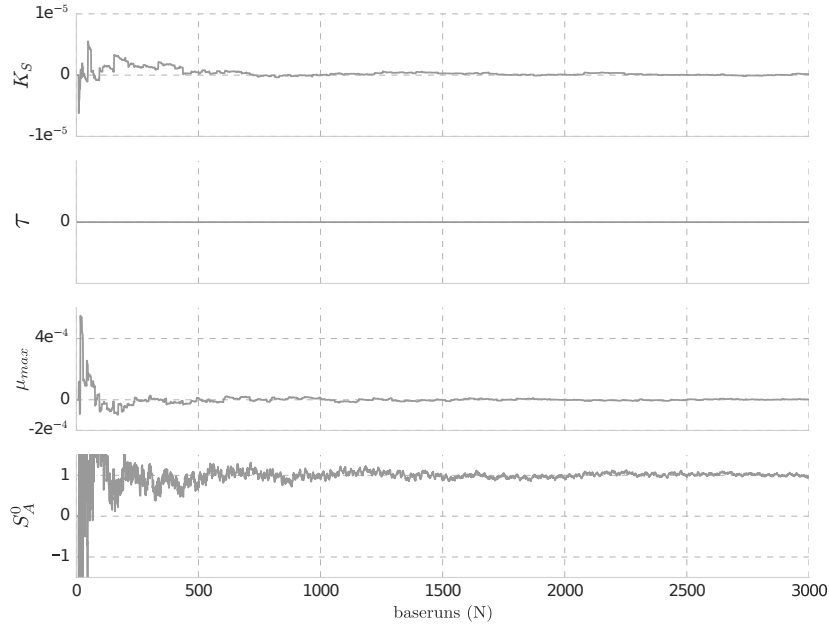


Figure 4.9: Progressive evolution of the estimated first order indices as function of the amount of base runs executed for the result provided in Figure 4.8a. For a small set of simulations, the estimated values are very mutable. For 3000 base runs (equivalent to 18000 simulations), the results have converged towards values which provide quantative information.

When the convergence is very slow, normalization of the metric values can improve the convergence of the sensitivity indices (Nossent and Bauwens, 2012a).

Extracting additional information

In order to properly investigate the differences or the similarities with the local sensitivity analysis, the derivation of the indices in function of time would be interesting as well. Without switching to another method, the available simulation outputs can also be reconsidered for alternative aggregation metrics. In order to make a comparison with the local sensitivity, the sensitivity of the average output of each minute of the simulation individually was calculated.

The model is providing the output in seconds, so the most convenient approach is to recalculate the output towards average minutes. Similar to the previous

section, the recalculated outputs can be used to derive the sensitivity indices, but the default plots would not be sufficient, so the output is summarized in the custom made Figure 4.10 for the first order effect and Figure 4.11 for the total order effect. For the total order effect, the sum is added to the plot as well. Values above 1 for the total order effect indicate more interaction effects in between the input factors.

In this case, only the first ten minutes are taken into account, since the direct application of later periods is influenced by the set of simulations for which the substrate is already consumed entirely. Adapting the possible parameter combinations would be an option to prevent this.

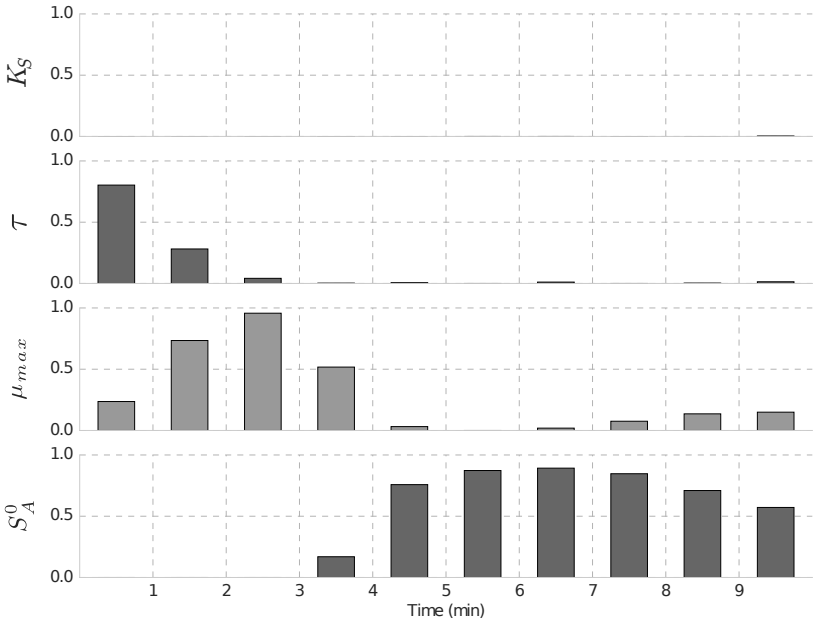


Figure 4.10: First order effects S_i of the average output for each simulated minute of the dissolved oxygen concentrations S_O

The relative importance of the substrate addition S_A^0 is large, both by its direct effect as well as in interaction with the parameters. This corresponds to previous literature mentioning the importance of the initial substrate on the ability to identify parameters (Grady et al., 1996). It seems rather counter-intuitive that the effect of an initial condition is mainly affecting the simulation in a later stage (after approximately 3 minutes). However, since each experiment starts from a

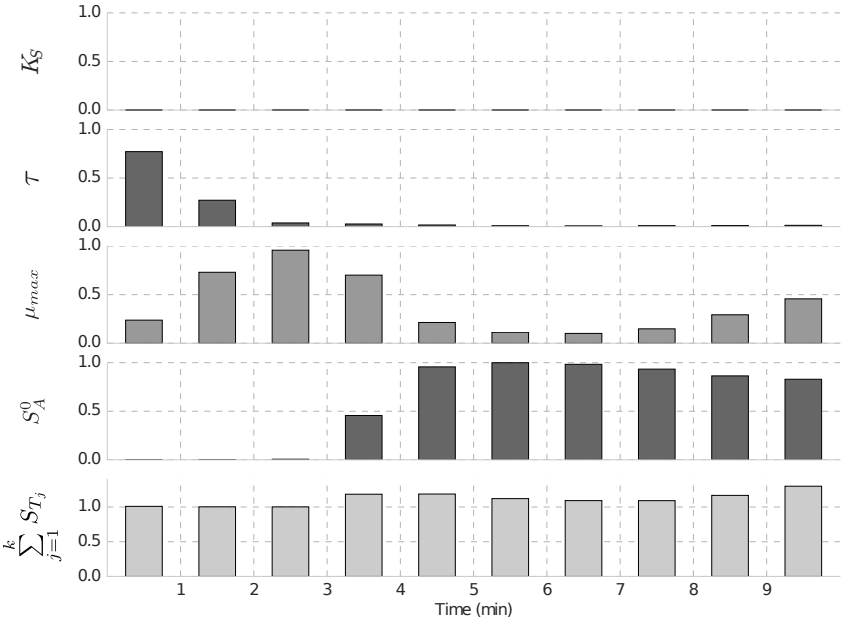


Figure 4.11: Total order effects S_{T_j} of the average output for each simulated minute of the dissolved oxygen concentrations S_O

similar initial condition of oxygen S_O^0, the substrate amount only affects the later stage. The added S_A^0 is more important after the first phase, since it defines the length of the period during which the biomass is fully active.

In general, the importance of K_S is very low, making it an option for parameter fixing when it would be part of a larger set of parameters (Figure 4.10 and Figure 4.11). The relative importance of parameter μ_{max} is lower in comparison to the initial substrate, but still more important than the other parameters, which is in line with the local sensitivity analysis. Interaction between the parameters increases after a few minutes as shown by the sum of the total sensitivity indices of Figure 4.11, as the combined effect of the parameters and initial substrate is affecting the model output. Overall, the interaction effects are limited and the identification of the parameters should be feasible with this set of parameters considered. This is in line with the conclusions of the structural identifiability analysis of Dochain et al. (1995) for a model without a lag-time, listing three identifiable combinations: $\mu_{\max} X_B (1-Y)/Y$, $(1-Y)S_A^0$ and $(1-Y)K_S$. Given that, in this

case the yield Y and the biomass X_B are assumed known, the identification of the parameters μ_{max}, K_S, τ and the initial substrate S_A^0 should be feasible.

The main influencing factor in the first minute is parameter τ. Hence, the global sensitivity analysis suggests that overall, the experimental condition has a major influence on the output and the parameter τ should be able to be estimated well over a larger range of different experimental conditions. Together with the information from the local sensitivity method, experimental conditions with a high S_A^0/X_B^0 ratio are preferred. Still, too large acetate concentrations could alter the physiological state of the biomass during the experiment, which should be avoided (Grady et al., 1996).

4.5 Model calibration

The focus is on the estimation of the parameters μ_{max}, K_S and τ, using the derived OUR_{ex} values of the single respirogram of Figure 4.2.

Different performance metrics can be used to compare the modelled output with the observations, some of which can be translated into an existing theoretical framework (section 3.4.2). In later chapters of the dissertation the focus is on informative performance metrics supporting the exploration of the model behaviour rather than using a specific theoretical framework. This section aims to illustrate the application of a specified likelihood function as performance metric in theoretical frameworks such as Maximum likelihood estimation and Bayesian approaches. The example is a translation of the work presented by VanderPlas (2014) and Foreman-Mackey et al. (2013) towards an ODE based model.

Consider the response surface plot in Figure 4.1, suggesting the interaction between the parameter μ_{max} and K_S. It represents the response surface of the SSE as a function of the parameter combinations. We can represent the SSE as a probability function as well, since it assumes a normal distribution for the residuals ($i = 1, \ldots, N$), i.e. independent and with zero mean. The likelihood function is constructed by taking the product of their normal distributions, similar to Equation 3.6:

$$P(\mathbf{y} \mid \mu_{max}, K_S, \tau) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \hat{y}_i}{\sigma}\right)^2\right] \tag{4.7}$$

The standard deviation σ is considered constant (homoscedastic) and is estimated as the standard deviation of the set of observations. $\hat{\mathbf{y}}$ represents the set of mod-

elled and $\mathbf{y}$ the set of observed oxygen uptake rates OUR_{ex}. For the implementation, the log-likelihood will be used, since likelihoods can be summed, whereas products of the probabilities of many data points tend to be very small.

To use a Bayesian approach, recall that the theorem of Bayes (Equation 3.10) states

$$P(\mu_{max}, K_S, \tau \mid \mathbf{y}) \propto P(\mathbf{y} \mid \mu_{max}, K_S, \tau) P(\mu_{max}, K_S, \tau) \tag{4.8}$$

of which we already defined the likelihood function $P(\mathbf{y} \mid \mu_{max}, K_S, \tau)$. Hence, in order to calculate the posterior function $P(\mu_{max}, K_S, \tau \mid \mathbf{y})$ the prior function $P(\mu_{max}, K_S, \tau)$ need to be decided on as well. Since no specific information is available (and to make the connection with the maximum likelihood estimation later on), uniform (so-called uninformative) priors will be used, with similar boundaries as those used in the global sensitivity analysis case (section 4.4).

Approximating the posterior by brute force would be largely inefficient, so the usage of an MCMC sampling method is appropriate. The emcee Python Package 3 Foreman-Mackey et al. (2013) is a lightweight pure-Python package which implements the Affine Invariant Ensemble MCMC method. The method provides improved convergence compared to classic MCMC sampling methods, mainly in the case of skewed distributions (Goodman and Weare, 2010).

> **Python Package 3** (emcee)**.**
> emcee is an MIT licensed pure-Python implementation of Goodman & Weare's Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler with a well written manual. emcee is being actively developed on GitHub.
> (http://dan.iel.fm/emcee/current/)

The sequence of samples is shown in Figure 4.12 for each of the parameters separately. After a period of burning-in, the samples provided by the Markov chain are exploring the posterior distribution and the samples after the burning in period can be used to construct one and two dimensional projection (histograms) of the posterior probability distributions of the involved parameters. The result is shown in Figure 4.13, which demonstrates all of the interactions (covariances) between the parameters and the marginalized distribution for each parameter independently. The density plot was made using the `corner` plot Python Module 4 (Foreman-Mackey et al., 2014).

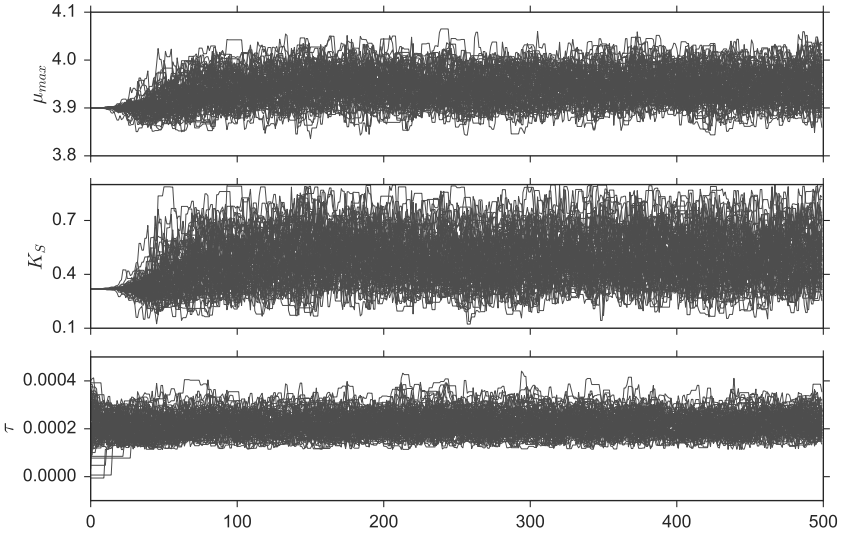


Figure 4.12: Progression of the MCMC sampler showing the individually taken samples while sampling the posterior parameter distributions after application of a Gaussian likelihood function and conditioned by the observations of Figure 4.2.

The limited amount of interactions observed in the global sensitivity analysis is also for this subset of parameters recognized in the resulting two-dimensional density plots. The optimization is performed on another data set as the local sensitivity analysis (section 4.3), but the modelled experiment of Cierkens et al. (2012) can be considered as a high S_A^0/X_B^0 ratio situation, for which the identification of parameter τ is represented here as well.

Notice the equivalence with the response surface information shown in Figure 4.1, focusing on parameters μ_{max} and K_S. However, that figure provides the information about the response surface (using SSE) of the parameters, whereas in the case of Figure 4.13 the graph represents the density of the posterior parameter distributions by the defined likelihood. Still, the equivalence between the result concerning the interaction effect between μ_{max} and K_S is important to notice.

This is further illustrated by expressing the same likelihood function of Equation 4.7 as an optimization problem, estimating the maximum likelihood. Since it is a pure optimization problem, the application of the scipy optimize Python Module 2 is appropriate. The fact that the scipy optimize function searches for a

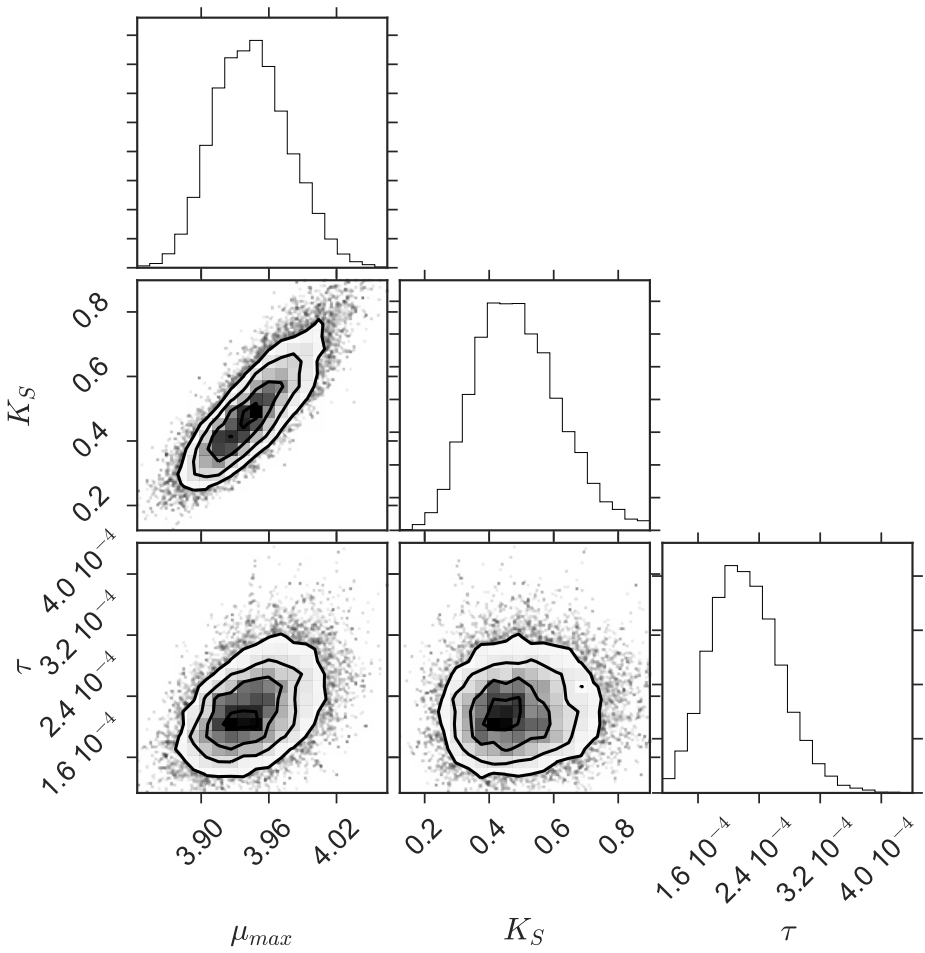


Figure 4.13: Corner plot of the samples constructing the posterior parameter distributions after application of a Gaussian likelihood function and conditioned by the observations of Figure 4.2.

minimum, while we want to maximize the likelihood can be solved by minimizing the negative log-likelihood. The application of the optimization gives: $\mu_{max} = 3.93$ d^{-1}, $K_S = 0.45$ $\mathrm{mg\,l}^{-1}$ and $\tau = 2.1 \times 10^{-4}$ d which agrees to the highest densities of the distributions found by the MCMC approach (see Figure 4.13). Hence, under the assumption of uniform priors, the Bayesian probability is maximized at precisely the same value as the ML result (MacKay, 2002; VanderPlas, 2014).

Python Module 4 (corner/triangle)**.**
Module to make an illustrative representation of one- and two-dimensional projections of samples in high dimensional spaces.

The module is built by Dan Foreman-Mackey and collaborators (see triangle.__contributors__ for the most up to date list). Licensed under the 2-clause BSD license.
(https://zenodo.org/record/11020)

4.6 Conclusion

The example case study in this chapter uses different approaches investigate the characteristics of a respirometric model making use of a set of experimental data. The properties of the model structure as well as the identification of the parameters under different experimental conditions are examined.

A local sensitivity analysis emphasises the importance of the experimental conditions to support the identification of the model parameters. It indicates that the addition of the time-lag in the model structure needs to be supported by experimental data for which the S_A^0/X_B^0 ratio is sufficiently high. A high ratio provides the ability to estimate the parameter τ of the time-lag model component.

The entire parameter space is taken into account by using a global sensitivity analysis. The relative importance of the added substrate on the total model output variability compared to the effect of the parameters is assessed. The importance of the chosen initial substrate concentration on the assessment of the model structure should be taken into account when performing lab experiments. Moreover, the degree of interaction between the considered parameters is relatively low, confirming the ability to identify the parameters μ_{max}, K_S and τ and the suitability of the proposed model structure.

Finally, the model parameters are estimated for the case of a high S_A^0/X_B^0 ratio. As expected from both the local and global sensitivity analysis, under these conditions the interaction effects are limited, leading to an identifiable region for the considered parameters.

These results are in line with earlier work focusing on the identifiability of the parameters μ_{max} and K_S for a more simplified respirometric model (Dochain et al.,

1995; Vanrolleghem et al., 1995). Moreover, it is shown that the addition of a time-lag component to capture the retardation of the biomass activity can be justified and parameter the time-lag parameter τ is practically identifiable as well under the used assumptions, when using a proper S_A^0/X_B^0 ratio.

At the same time, the analysis illustrates the central position the metric selection takes. By applying the same likelihood function within the scope of an ML estimation (optimization) and a Bayesian approach (sampling), it illustrates how the decision of a metric and its involved assumptions are not bound to a single method, but can be reused by different algorithms. Both aggregated as time dependent metrics can be used to derive sensitivity indices, always in function of the research objective.

CHAPTER 5

Sensitivity Analysis methods

Parts redrafted and compiled from

Van Hoey, S., Seuntjens, P., van der Kwast, J., de Kok, J.-L., Engelen, G., and Nopens, I. (2011). Flexible framework for diagnosing alternative model structures through sensitivity and uncertainty analysis. In Chan, F., Marinova, D., and Anderssen, R. S., editors, *MODSIM2011, 19th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand*, pages 3924–3930. Modelling and Simulation Society of Australia and New Zealand (MSSANZ)

5.1 Introduction

In general, SA focuses on the response of a model output to changes of model input factors. The aim is to get insight in how the changes of the output can be attributed to the variations of the inputs. Inputs are not limited to model parameters alone, but can be any input factor that drives variation of the model output (Saltelli et al., 2008). Similarly, model output can be a state variable itself on any given time step, but also any aggregated or performance metric. The choice of the input factor and output metric should always be directly linked to the research question.

Methods for SA play a central role in the evaluation of model structures and support the model diagnostic process (Wagener and Kollat, 2007). Different techniques are available in literature and the applicability depends on the model characteristics, the dimensionality of the problem and the available computational time

(Tang et al., 2007b; Yang, 2011). The application is not always straightforward in the case of non-linear and high-dimensional models as faced in environmental modelling. This leads to a sprawl of available methods, characterized by different assumptions, changing conditions of application and various code implementations.

Whereas the application of SA is well-recognised in the environmental modelling community, the execution and reporting of sensitivity analysis is sometimes hampered due to the lack of well-documented implementations. Tools are needed to facilitate the usage of SA techniques also by non-specialist users, as well as to provide guidelines on GSA application (Pianosi et al., 2015). Moreover, code documentation is regularly ignored, driven by the perception that the code is not written for others to use Petre and Wilson (2014).

To overcome the lack of code documentation, facilitate re-use and provide transparency of existing implementations, the methods implemented within the scope of this dissertation were collected in a dedicated Python package, called pystran. The provided code is mainly to provide transparency in the implementation and is not a finished software product.

For the source code documentation, the reader is referred to the online documentation[1]. The aim of this chapter is to provide the theoretical background on the SA methods as they were implemented and used within the scope of this dissertation. Next to a description of the individual methods, a flowchart to provide guidance to the modeller in the selection of a specific SA method is proposed in the last section.

Python Package 4 (pystran)**.**
The pystran package collects a set of methods for to perform sensitivity analysis with a specific focus on model evaluation. Following the metric oriented approach described in section 3.2.2, the pystran package supports the easy linkage with a set of model performance metrics. The package provides an open and extensible implementation, written in the python programming language. Several plot functions are built-in to facilitate the execution and interpretation of the implemented methods. The open source licence of the pystran package provides the ability for other users to further develop and improve the implementation.
(https://github.com/stijnvanhoey/pystran)

[1] http://stijnvanhoey.github.io/pystran/

Readers who are familiar with these SA techniques can safely skip this chapter, as it is mainly a reference on the theoretical background and guidance on the interpretation of the methods. A subset of the methods is used in the other chapters, for which the reader can always return later when more background information is required on a particular method.

5.2 Sensitivity analysis: general remarks

A complete overview of existing methods for sensitivity analysis is not the purpose of this chapter. In the remainder of this chapter focus is given to those methods that are either implemented or applied in the other chapters. In-depth reading material is provided by Saltelli et al. (2008), giving a more complete overview of existing methods. Moreover, additional reviews and comparative studies can be found in literature (Frey and Patil, 2002; Tang et al., 2007b; Lilburne and Tarantola, 2009; Mishra, 2009; Gan et al., 2014; Vanrolleghem et al., 2015).

Different rationales to perform a sensitivity analysis do exist, which depend on the field of interest and the application.

- The identification of the most influential factors can support uncertainty analysis. The factors with the most influence should be focussed on to increase robustness, since their uncertainty will have a major influence on the model uncertainty if their uncertainty is large. Notwithstanding the direct link between sensitivity analysis and uncertainty analysis, it is important to understand that sensitivity analysis only tells something about the potential influence on the uncertainty and does not provide any predictive statement about the uncertainty itself.
- To facilitate model calibration, i.e. by identifying critical regions in the parameter space. Hence, focus is given to the model parameters with most influence, which is also referred to as factor prioritization (Saltelli et al., 2008).
- Opposite to factor prioritization, the identification and fixing of non-influential parameters (factor fixing) reduces the dimensionality of the problem. Furthermore, the removal of redundant parts leads to simplification of the model.
- Sensitivity analysis is of major importance for model evaluation (section 2.3.3). Input interactions can be assessed and the identifiability of individual inputs can be checked.

A general division between local and global methods for SA can be made. Whereas local methods (see section 4.3) focus on a specific location in the parameter space, the global methods consider the whole variation range of the factors. In the case of global sensitivity analysis, most methods are based on a sampling strategy of an assumed parameter distribution (see section 3.5). However, many global sensitivity methods use a specialised sampling strategy in order to better support their analysis. As such, these sampling schemes are usually introduced together with the method itself, but basically extend the possible set of sampling schemes. They could be used as sampling procedures for other applications as well.

It is noteworthy that the methods described in the next sections use input factors as the input of the sensitivity analysis. Model parameters are only a subset of the total set of possible factors used in an SA method. Hence, $\boldsymbol{\theta}$ (and $\boldsymbol{X}$ in the case of Variance based methods) represents the input factor, which can be a parameter as well as another input.

In line with chapter 3, all of the methods do act on a chosen aggregated or performance metric. Besides the sensitivity towards an aggregation metric (variable of interest), also the sensitivity towards performance metrics can be assessed or both can be used as variable of interest. The decision of a metric does not restrict the modeller to a single method, since the methods described in the next sections are generally not restricted by theoretical considerations on the chosen metric.

Before explaining the methods themselves, some information about the accompanying schemes (Figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6 and 5.8) is given. For each of the methods, a representative scheme is provided, summarizing the method schematically in four sub-figures. A similar concept will be used for each of the methods to provide a step-wise overview of the different methods, which illustrate also their similarities. For each of the visualisations, sub-figure (a) focuses on the sampling strategy linked to the methodology. For some methods, this is a direct sampling of the input factor space, whereas for other methods this is a trajectory sampling. The small axis inside sub-figure (a) represents a random sample from an input factor distribution, for which a uniform distributions with range $[0 - 1]$ is shown. It should be noted that this can also be a sample from any non-uniform distribution, as discussed earlier in section 3.5.

Sub-figures (b) and (c) illustrate the translation of the model outcome into the sensitivity indices used for that particular technique. In the last sub-figure (d), a summarizing representation of the results of the analysis is given on which the interpretation of the analysis is typically based. It is important to understand that the model output metric $\hat{y}_i$ in the figures can be any kind of aggregation or performance metric (e.g. a single time step of the model output, an average value of

the output, a performance metric taking into account available measurements...). In the visualisations, $\hat{y}$ refers to the model output as a function of time, whereas $\hat{y}_i$ can be any metric of interest, but for some models or when focusing on single time steps both are actually the same.

5.3 Morris Elementary Effects (EE) screening approach

Screening methods can be used to isolate that set of factors that has the strongest effect on the output variability with relatively few model evaluations. This makes it an appealing technique for computationally expensive models investigated for a large set of input factors (Saltelli et al., 2008; Morris, 1991). It also makes them appropriate as an initial stage preliminary tool (to reduce the dimensionality of the problem), before a more detailed analysis is performed (Campolongo et al., 2011). Hence, it mainly provides a qualitative assessment to rank the input factors in their order of importance and to make statements about being *more* and *less* sensitive. The Morris method is particularly well-suited when the number of input factors is high and/or the model is expensive to compute, providing a very good compromise between accuracy and efficiency (Campolongo et al., 2007). As a screening tool, it is able to screen the most and least influential parameters for a highly parameterized watershed model with 300 times fewer model evaluations than variance based methods (Herman et al., 2013a). Still, Nossent et al. (2013) and Vanrolleghem et al. (2015) illustrate the importance of a proper convergence assessment to prevent the incorrect elimination of influential factors.

5.3.1 Elementary Effects (EE) based sensitivity metric

The Elementary Effect (EE) global screening method by Morris (1991) is a One factor At a Time (OAT) based method that is based on the calculation of so-called Elementary Effects (EEs). These EEs are similar in nature to the local SA finite difference approximation as defined in a local sensitivity analysis (Equation 4.6). Assume an application on a set of k different factors $\boldsymbol{\theta}$ of a model defined by Equation 2.1. The EE of factor θ_j towards a variable of interest $\hat{y}_i$ (any kind of aggregation on the model output) is defined as follows:

$$EE_{\theta_j} = \frac{\hat{y}_i([\theta_1, \ldots, \theta_{j-1}, \theta_j + \Delta_{EE}, \ldots, \theta_k]) - \hat{y}_i(\boldsymbol{\theta})}{\Delta_{EE}} \tag{5.1}$$

with $\hat{y}_i([\theta_1, \ldots, \theta_{j-1}, \theta_j + \Delta_{EE}, \ldots, \theta_k])$ the value of $\hat{y}_i$ when Δ_{EE} is added to the value of factor θ_j and could be rewritten as $\hat{y}_i(\boldsymbol{\theta} + \Delta_{EE})$ similar to Equation 4.6. The difference with the local procedure is the usage of Δ_{EE}, which is a predetermined multiple of $1/(p-1)$ where p is the number of levels of the design. p corresponds to the number of levels the regular k-dimensional grid of factors is discretized in. Hence, the EE can be calculated for any θ_j between 0 and $1 - \Delta_{EE}$ where $\theta_j \in \{0, 1/(p-1), 2/(p-1), \ldots, 1\}$. When other distributions are assumed for θ_j, the values sampled in the interval $[0-1]$, should be converted by using the inverse method (section 3.5.1) and used as such by the model. The influence of θ_j is then evaluated by computing several EEs and assess the effect.

The evaluation of the sensitivity is done based on the originally proposed sensitivity measures (Morris, 1991), namely the mean μ_j and the standard deviation σ_j of the calculated EEs and also on the mean of the absolute values of the EE, μ_j^*, as recommended by Campolongo et al. (2007). The latter prevents that EEs with different signs are cancelling each other out. In most applications, the combined analysis of the three indices is recommended to extract the maximum amount of information (Saltelli et al., 2008).

Furthermore, μ_j^* provides a good proxy to the Sobol total sensitivity index S_T (Saltelli et al., 2008; Yang, 2011) due to its effective screening capability. The total effect S_{T_j} of factor θ_j, corresponds to the effect of the individual factor in combination with all the interactions of this factor with the other factor. The relative variance it represents when all factors but the jth factor are fixed, is used to check for potential factor fixing of factors. The latter action means that one sets certain factors to fixed values when they have an $S_{T_j} = 0$, i.e. they do not influence the output variability of the (aggregated) variable at all.

It is important to understand that Morris provides a screening of the input factor space, which results in qualitative results rather than quantitative estimates of the factor influence. This provides interpretations in terms of less and more influential factors (ranking). To summarize the important aspects of the interpretation of the sensitivity indices:

- A low value for μ_j^* indicates that the factor has a limited influence on the (variance of the) response variable
- A high value for σ_j highlights the interaction between different factors and/or the non-linearity of the model
- Comparison of μ_j with μ_j^* provides information on the sign of the influence of the effect of the factor

5.3.2 Sampling strategy

To compute r different EEs for each of the k factors, a total of $2rk$ model simulations would be needed, with a random sampling step for each of the r EEs for which the methods in section 3.5 can be applied. Morris (1991) used a sampling strategy that belongs to the class of OAT (one factor at a time) designs, but designed it efficiently by making use of r trajectories of $(k+1)$ points in the input space, each providing k elementary effects, hence with a total of only $r(k+1)$ simulations. For each trajectory, the input space dimensions are one by one traversed starting from a randomly sampled base point $\boldsymbol{\theta}^*$ (for which the sampling techniques described in section 3.5 can be used). A detailed description of the sampling strategy is provided in Morris (1991) and well-explained by Saltelli et al. (2008).

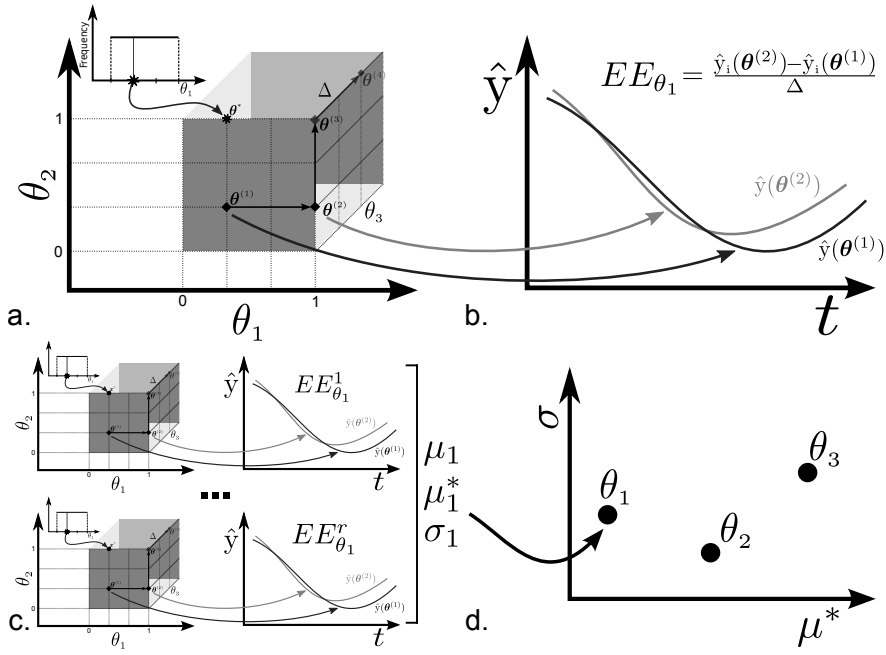


Figure 5.1: Overview of the Morris method based on the derivation of EEs. For each trajectory, a random sample of the input factor space is taken, after which the different dimensions are traversed, each by a step Δ_{EE} (**a**). Based on a combination of two consecutive runs, an EE can be calculated for a single input factor (**b**). By running multiple trajectories, the summarizing sensitivity indices μ_j, μ_j^* and σ_j can be calculated (**c**) and graphically expressed in a (μ_j^*, σ_j)-plane (**d**).

Figure 5.1 summarizes the concept of the Morris method applied to a set of input factors $\boldsymbol{\theta}$. Figure 5.1a illustrates the sampling of a single trajectory over a three dimensional input space. A trajectory starts with the sampling of a base point $\boldsymbol{\theta}^*$ in one of the p levels for each of the factors (represented by the small axis for factor $\boldsymbol{\theta}_1$). Starting from this base point, a consecutive set of parameter sets is sampled, each time changing a single factor with a value Δ_{EE}. For a three dimensional set of input factors, the model needs to be simulated with the parameter sets $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\theta}^3$ and $\boldsymbol{\theta}^{(4)}$. The resulting model outputs for the two runs $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are shown in Figure 5.1b (for this example as a function of time, but this can be any metric). These two outputs can be used to calculate an EE for factor θ_1 (EE_{θ_1}). Hence, a single EE for each input factor is calculated by running a single trajectory. In order to get a global sensitivity metric, a set of r trajectories is used to calculate r EEs, represented in Figure 5.1c. The set of EEs is summarized in the set of sensitivity indices μ_j, μ_j^* and σ_j (in the figure only shown for factor θ_1). The graphical representation of the (μ_j^*, σ_j)-plane provides insight about the factor importance (μ_j^*) and the interaction effects (σ_j), which is shown in Figure 5.1d. This is the result of the analysis that can be used for evaluation.

A further improvement of the sampling strategy has been proposed by Campolongo et al. (2007). It aims to improve the scanning of the input domain without increasing the number of model evaluations. The method selects a subset of trajectories with the highest spread, out of an initially large set of generated trajectories, by maximizing the distance between the pairs of trajectories (Campolongo et al., 2007; Saltelli et al., 2008). The distance between a pair of trajectories m and l, d_{ml}, is defined as is the sum of the geometric distances between all the couples of points of the two fixed trajectories (Euclidean distance):

$$d_{\mathrm{ml}} = \begin{cases} \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \sqrt{\sum_{z=1}^{k} \left[\theta_i^{(m)}(z) - \theta_j^{(l)}(z)\right]^2} & m \neq l \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $\theta_i^{(m)}(z)$ indicates the zth coordinate of the ith point of the mth trajectory and $\theta_j^{(l)}(z)$ indicates the zth coordinate of the jth point of the lth trajectory. Each trajectory is composed of k factors and the base point $\boldsymbol{\theta}^*$. Hence, the Euclidean distance needs to be summed for all combinations of $k+1$ points. The best r trajectories out of M are selected by maximising the distance d_{ml} among them using any optimization scheme. In the pystran Python Package 4, a brute force approach is chosen by comparing all possible combinations. The usage of the preliminary optimization procedure will cancel out the advantages of an improved

sampling approach (see next section) of the base point $\boldsymbol{x}^*$ (Campolongo et al., 2007).

5.3.3 Working with groups

The EE method can also be applied to work with groups of input factors instead of single factor values, which is most useful in the case of very large dimensional problems. It allows for the reduction of the number of simulations, at the cost of not obtaining information about the relative strength of the inputs that are merged in a group (Campolongo et al., 2007; Saltelli et al., 2008).

The usage of μ_j is not possible in the case of grouped input factors, since two factors within a single group could have opposite influence on the response variable. Hence, the interpretation will be based on μ_j^* instead. The sampling scheme needs to be adapted as well, as described by Campolongo et al. (2007).

The technique of groups is not applied in the remainder of this dissertation apart from the flow chart proposed in section 5.10. However, it was implemented and tested within the scope of the pystran Python Package 4. Hence, for further information about the functionalities and their handling, the reader is referred to the documentation of Python Package 4.

5.4 Global OAT sensitivity analysis

An alternative method of the well-known Morris screening method has been proposed by van Griensven et al. (2006), aiming to combine the robustness of an improved sampling scheme with the functionality of an OAT approach. They provide a direct translation of the local SA methodology towards a global technique, taking r Latin Hypercube (LH) samples in the parameter space, and then varying each sampled point k times by changing each of the k factors one at a time, as is done in the OAT design. In short, the method executes a local SA in r different points in the parameter space, resulting in a trajectory in each point (called *loops* in van Griensven et al. (2006)). Within each of the trajectories, the so-called *partial effect* PE_{θ_j} (similar to the EE of Morris method) of a factor θ_j towards a variable of interest $\hat{y}_i$ (any kind of aggregation or performance metric) is calculated as:

$$PE_{\theta_j} = \left| \frac{100 \cdot \left(\frac{\hat{y}_i([\theta_1,\ldots,\theta_{j-1},\theta_j\cdot(1+f_j),\ldots,\theta_k]) - \hat{y}_i(\boldsymbol{\theta})}{\hat{y}_i([\theta_1,\ldots,\theta_{j-1},\theta_j\cdot(1+f_j),\ldots,\theta_k]) + \hat{y}_i(\boldsymbol{\theta})} \right)}{f_j} \right| \tag{5.3}$$

with f_j the fraction by which factor θ_j is changed (a predefined constant) and with $\hat{y}_i([\theta_1, \ldots, \theta_{j-1}, \theta_j(1+f_j), \ldots, \theta_k])$ the value of $\hat{y}_i$ when $(1+f_j)$ is multiplied with the value of factor θ_j. Hence, it could be rewritten as $\hat{y}_i(\boldsymbol{\theta} \cdot (1+f_j))$ as well (Equation 4.6). Similar to the Morris approach, the influence of a factor θ_j is calculated by averaging these partial effects of each loop for r trajectories, also leading to $r(k+1)$ required simulations. Since the aim is to provide qualitative information about influence of the factors, mostly the rank of each factor will be communicated.

The procedure of the global OAT approach is summarized in Figure 5.2. The visualisation is very similar to Figure 5.1. Figure 5.2a represents a single trajectory and starts from the random sampling of a base point for each of the factors (see small axis for θ_1). In contrast to Figure 5.1, the sampling is not based on a fixed set of levels, but can be any sampled value in the factor space. Furthermore, the step between two parameter combinations is defined by the relative factor f_j instead of Δ. The two simulations $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are used to calculate the partial effect PE of input factor θ_1, similar to the Morris approach (Figure 5.2b). Hence, a PE for each input factor is calculated by running a single trajectory. Figure 5.2c represents the usage of r trajectories and the influence of the factor is estimated by the average of the individual partial effects. The average values $\bar{PE}_{\theta_i}$ are represented by sorting their values and checking the relative importance of each of the factors, as shown in Figure 5.2d. Alternatively, when used for multiple metrics, a table representation is used as well, listing the ranks for each of the outputs.

Within the scope of the pystran implementation, a decoupling of the basic elements enabled a further generalisation in the implementation compared to van Griensven et al. (2006):

- Sampling of the input factor distributions can be performed by the methods described in section 3.5, for which LH is just one option. Hence, the method is here referenced as global OAT.
- The fraction by which factor θ_j is changed within each step of a trajectory, is the same as the finite difference approximation of a local sensitivity analysis (section 4.3)
- Apart from the partial effect defined by Equation 5.3, also the absolute and total relative sensitivity from the local sensitivity method described in section 4.3 can be evaluated within each trajectory.

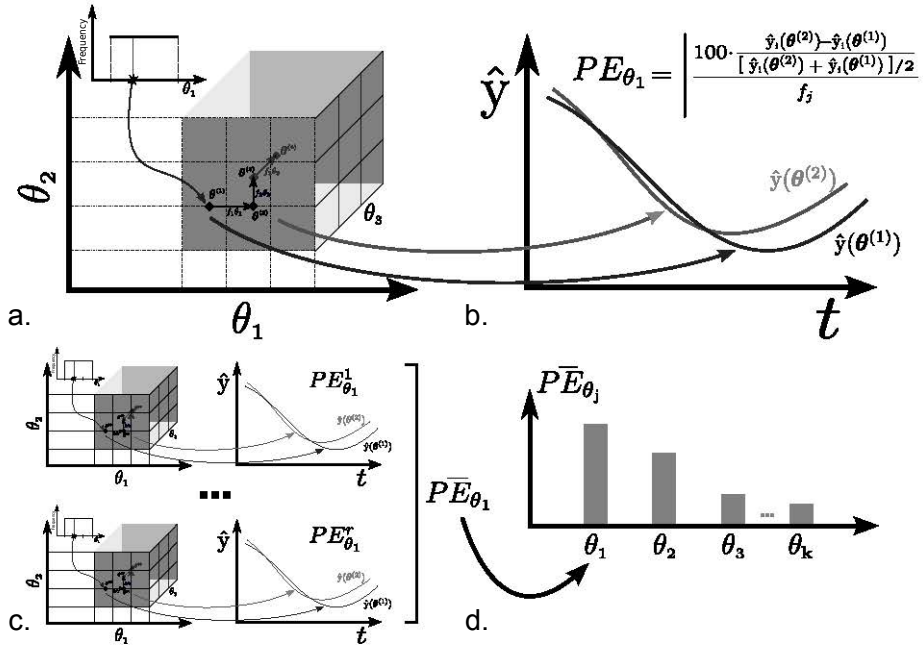


Figure 5.2: Overview of the Global OAT method, which is very similar to the Morris approach of Figure 5.1. For each trajectory, a random sample is taken from the factor distributions which is used as a starting point from which k other simulations are performed with a stepsize $f_j\theta_j$(**a**). For each couple of consecutive simulations, the partial effect of an input factor θ_j can be evaluated for the variable of interest $\hat{y}_i$ (**b**). By performing a sufficient set of r trajectories (**c**), the partial effects can be summarized by their mean value to provide information about the sensitivity of the individual input factors θ_j (**d**).

The method will not be applied in the remainder of the dissertation as it was only used as a reference towards the original Morris screening approach. More information is provided in the online manual of the pystran Python Package 4.

5.5 Standardised Regression Coefficients

A sensitivity metric of a response variable can be obtained using an emulator (also known as metamodel or surrogate model), which is any (more simple) mathematical function that approximates the relation between the considered input parameters and the response variable. The usage of emulator models is a separate

scientific discipline by applying any kind of machine learning (data based) techniques available to mimic the process based model one is working with (Saltelli et al., 2008). The application of machine learning techniques is not considered in this dissertation, the aim is to derive information about the process model itself. However, the most straightforward approach, i.e. the usage of a multiple linear regression model for which the regression coefficients provide an estimate of the sensitivity, is supported by the pystran Python Package 4 and will be shortly introduced. The method is well-known in the waste water modelling community (Vanrolleghem et al., 2015).

The regression is based on a set of N simulations by sampling from the assumed parameter distributions using any sampling strategy favoured (section 3.5). The linear regression model will use these samples to approximate the response variable y_i with $i = 1, \ldots, N$ (y is any aggregated metric of the process model, as it was defined $\hat{y}_\mathrm{i}$ in Equation 2.1, but now considered as the available 'data' for the regression model) by the set of input parameters $\boldsymbol{\theta}$ as follows:

$$y_\mathrm{i} = \beta_0 + \sum_{j=1}^{k} \beta_\mathrm{j} \theta_\mathrm{ij} + \varepsilon_\mathrm{i} \tag{5.4}$$

where β_j are regression coefficients to be determined and ε_i is the error between the process based model and the regression model due to the approximation. Under the assumption of Gaussian errors (i.e. the difference between the process based model and the regression model), the regression coefficients can be computed using the OLS approach, as it was implemented in the pystran Python Package 4.

The regression coefficients β_j with $j = 1, \ldots, k$, define the linear relationship between the parameters and the response variable. The sign of β_j defines the relation between the parameter θ_j and the response variable to be proportional (positive coefficient) or inverse (negative coefficient). The coefficients are dependent on the units in which $\boldsymbol{\theta}$ and y_i are expressed. Hence, the sensitivity metric (S_i) used for comparison is the standardized regression coefficient (SRC):

$$\mathrm{SRC} = \beta_\mathrm{j} \frac{\hat{s}_{\theta_j}}{\hat{s}_y}$$

where

$$\hat{s}_{\theta_j} = \left[\sum_{i=1}^{N} \frac{(\theta_\mathrm{ij} - \bar{\theta}_\mathrm{j})^2}{N-1} \right]^{1/2} \qquad \hat{s}_y = \left[\sum_{i=1}^{N} \frac{(y_\mathrm{i} - \bar{y})^2}{N-1} \right]^{1/2}$$

For the practical implementation, the parameter values and variables are normalized to mean zero and standard deviation one before applying regression, by

which the resulting regression coefficients are standardized. The regression based approach is summarized in Figure 5.3 for a set of θ_j input factors. Figure 5.3a illustrates the sampling procedure leading to a set of N model simulations. In contrast to the previous methods, the sampling strategy is not based on a trajectory, but can be directly performed by sampling N parameter sets from the input factor distributions. Each dot in Figure 5.3b represent the variable of interest $\hat{y}_i$ resulting from a simulation. A regression model (visualised as a grey plane in the figure) is estimated and the standardised regression coefficients represent the influence of the input factors. The latter is illustrated in Figure 5.3c as well, illustrating the partial effect of two factors on the chosen metric. Visualisation is mostly done in a bar chart such as in Figure 5.3d, sorting the values and making a clear distinction between positive and negative effects.

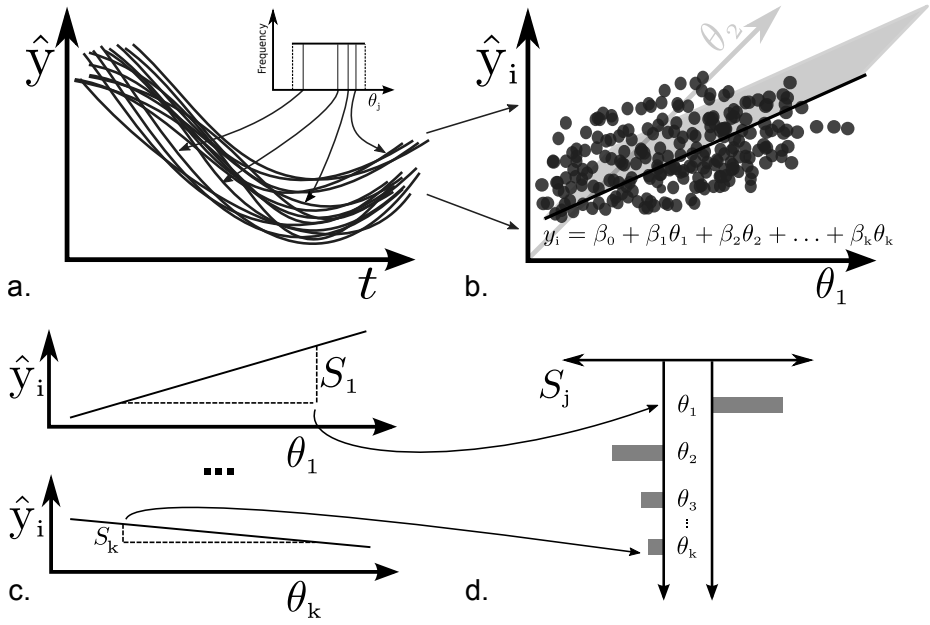


Figure 5.3: Overview of a regression based SA as implemented in the pystran Python Package 4. Similar to other methods, a sampling srategy is used to perform a number of simulations which are translated into an aggregated response variable $\hat{y}_i$ (**a**). A multivariate linear regression model is fitted (**b**) and the SRC coefficients define the influence of the corresponding input factor on the response variable selected (**c**). The results of a regression based approach is commonly done in a bar chart (**d**).

The applicability of the linear regression approximation needs to be evaluated, since the resulting SRC sensitivity metrics are only as good as the regression model is performing. Often, the coefficient of determination, R^2, associated with the linear regression is used to evaluate the appropriateness of the regression coefficients (Saltelli and Bolado, 1998). A value of 0.7 is generally used for acceptance of the linear model (Benedetti et al., 2012). Furthermore, the Variance Inflation Factor (VIF) can be used to check for collinearity. Large values (a threshold of 10) denote multicollinearity problems (Neter et al., 1996).

When the approximation of a linear model is not appropriate, the usage of a rank transformation can still be used for non-linear but monotone relations (Sieber and Uhlenbrook, 2005). Instead of the absolute values, the respective ranks are used to perform the regression and the resulting coefficients are called standardized rank regression coefficients (SRRCs). However, since the rank transformation modifies the model under analysis, the resulting coefficients can only be interpreted qualitatively (Saltelli and Bolado, 1998; Sieber and Uhlenbrook, 2005). The pystran Python Package 4 automatically provides both the SRC and SRRC coefficients.

Finally, a prerequisite for using SRCs as a sensitivity metric is the absence of parameter interactions. Otherwise, the resulting sensitivity will be dependent on interaction effects as well. In those cases, the usage of partial correlation coefficients (PCC) is more appropriate (Helton et al., 2006), but the latter is not implemented in the pystran Python Package 4.

It is clear that the usability of a regressed SA is rather limited due to the non-linear nature of most environmental models. Moreover, the lack of identifiability that is a central point within the evaluation of model structures is not compatible with the regression-based approach, which makes it unused in the remainder of the dissertation. However, since it is based on any set of MC simulations, it comes without any extra computational cost during the application of other SA methods. Hence, the application can be used in the exploration phase and supports the modeller in the learning process.

5.6 Variance based Sensitivity Analysis

5.6.1 Variance based methods

The main idea of the variance-based methods for SA is to quantify the amount of variance that each input factor θ_j contributes to the unconditional variance $V(\hat{y}_i)$ of the variable of interest $\hat{y}_i$. To align with the common notation of variance based SA (and probabilistic random variables), for the remainder of this section, Y will be used as the output response variable of interest $\hat{y}_i$ and $\boldsymbol{X}$ as the vector of input factors (in other sections defined as $\boldsymbol{\theta}$). In a similar fashion the unconditional variance of the output is $V(Y)$.

Hence, the aim is to rank the input factors according to the remaining variance taken over $\boldsymbol{X_{\sim j}}$ when factor X_j would be fixed to its true value x_j^*. The resulting conditional variance of Y is expressed as $V(Y|X_i = x_j^*)$ and is obtained by taking the variance over all factors except of X_j.

Normally, we do not know the true value x_j^* for each of the input factors X_j. Hence, instead of the real value, the average of the conditional variance for all possible values of X_j is used. This expectation value over the whole distribution of input X_j is defined as $E[V(Y|X_j)]$. Based on the unconditional variance of the output, $V(Y)$, the defined average and by using the following property of the variance:

$$V(Y) = V(E[Y|X_j]) + E[V(Y|X_j)] \tag{5.5}$$

the variance of the conditional expectation $V_j = V(E[Y|X_j])$ is obtained. This measure is also called the main effect, which is used as a sensitivity metric of the importance of an input factor X_j on the variance of Y. By normalizing the main effect by the unconditional variance of the output $V(Y)$ the first-order sensitivity index S_j is obtained:

$$S_j = \frac{V(E[Y|X_j])}{V(Y)} \tag{5.6}$$

The *first-order sensitivity index* S_j is mainly useful to identify the most important input factors (factor prioritization) and is a scaled value between 0 and 1. When dealing with additive models without interaction effects, the first-order indices of all input factors will explain the variance of the output. However, in the case of interaction effects, the sum of the first-order indices will be lower than 1 and

the remaining variance needs to be described by higher order interaction effects between different input factors.

The interaction effect in between two orthogonal (i.e. the attribution of the variance of each factor independently is possible) input factors X_j and X_l on the output Y can be expressed in terms of conditional variances as follows:

$$V_{jl} = V(E[Y|X_j, X_l]) - V(E[Y|X_j]) - V(E[Y|X_l]) \tag{5.7}$$

where $V(E[Y|X_j, X_l])$ measures the joint effect of the pair X_j and X_l. The joint effect of them minus the first order effects of the same factors, V_{jl} is called the second-order effect. Similar, higher-order effects can be computed. So, the variance of the third-order effect between the three orthogonal factors X_j, X_l and X_m would be:

$$V_{jlm} = V(E[Y|X_j, X_l, X_m]) - V_{jl} - V_{lm} - V_{jm} - V_j - V_l - V_m \tag{5.8}$$

For non-linear models the sum of all first order indices can be very low. Since non-linear models are common in environmental studies, the combined contribution from the first-order index in combination with all higher order interaction effects enables to assess the total effect of an input factor on the response variable. This sum of all the order effects that a factor accounts for is called the total-order effect and the *total sensitivity index* S_{T_j} is the sum of all indices relating to input factor X_j. The total sensitivity index can support the identification of input factors with limited overall influence on the output variance. A very low value of S_{T_j} indicates a minor effect of input factor X_j. Hence, their value can be fixed (factor fixing) or provide an indication for model reduction.

Figure 5.4 provides a visual overview of the variance based approach, where again a large set of simulations based on the random sampling of the input factor space and the associated model simulations are the start of the analysis, as shown in Figure 5.4a. Typically, a quasi-random sampling approach is used, which is based on a sampling of the input factor distributions. Similar to Figure 5.3, each black dot in Figure 5.4b represents the output of the variable of interest $\hat{y}_i$. The dots are vertically divided in narrow bands, and within each band the conditional mean $E[Y|X_j]$ is represented by a single grey dot. The variance of the grey dots, $V(E[Y|X_j])$ is used to estimate the first order sensitivity index of the factor. A higher number of bands and individual samples will improve the estimate of the sensitivity indices. Figure 5.4c illustrates how the influence of each individual factor and the interaction effects contribute to the total variance on the output $V(Y)$. The first order effect S_j is the variance by the factor itself, whereas the total sensitivity index S_{T_j} combines the variance provided by a factor and all the interactions of this factor with the other factors (different black arrows). The representation of

both S_j and S_{T_j} is typically done by bar charts or as tabular values, as illustrated in Figure 5.4d.

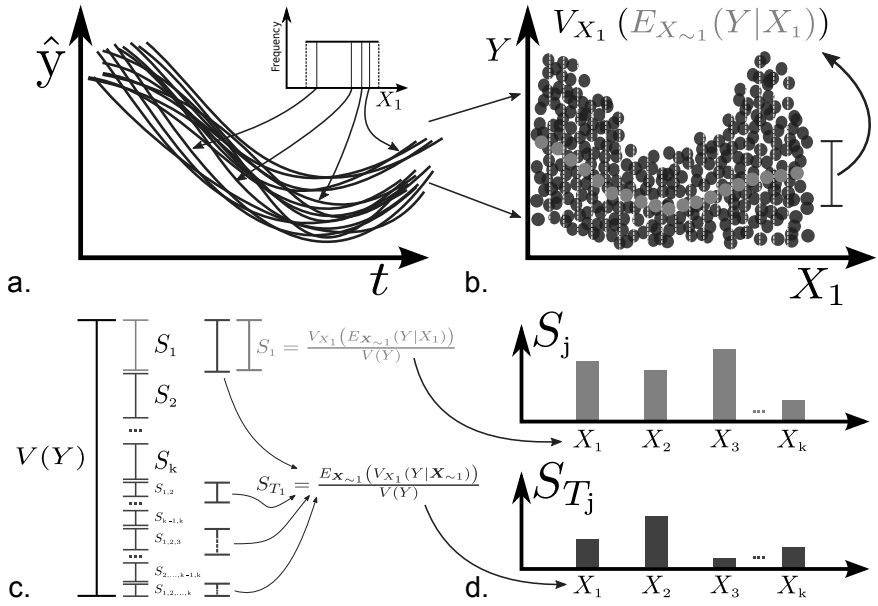


Figure 5.4: Overview of a variance based SA method, considering a set of simulations by random sampling the input space similar (**a**). The derivation of the variance V_j is done by calculating the variance of the conditional mean, which is the mean for a fixed value of X_j, represented by a narrow range (**b**). The first order S_j and total order S_{T_j} sensitivity index are describing that part of the variance respectively provoked directly by the input factor or in combination with other input factors (**c**). Communication can be done by using bar-charts (**d**) or as tabular values. Note the usage of X_j here to define input factors instead of θ_j and Y to define the metric of interest to agree with the common notation of variance based methods in literature.

The computation of all order-effects to calculate the S_{T_j} for each of the input factors X_j by brute force would result in the necessity of evaluating $2^k - 1$ different terms. Consider Figure 5.4b, which focuses on the derivation of a single term $V_j = V(E[Y|X_j])$. Assume 1000 simulations with a fixed value of X_j would be performed to get an estimate of the conditional mean $E[Y|X_j]$ (single grey dot within each narrow band) and this procedure would be performed for 1000 different values of X_j (thousand narrow bands with each a grey dot), the required set of

simulations for this single term would be 10^6. This makes the brute force approach infeasible for higher-dimensional models.

Different MC based methods exist to overcome this dimensionality problem and provide an approximation of the first, total and higher order sensitivity indices of k input factors. Homma and Saltelli (1996) illustrated how the total number of terms that need to be evaluated to have a good representation can be reduced to $2k$. Different approximating methods have been developed, such as the Fourier Amplitude Sensitivity Test (FAST) and Extended Fourier Amplitude Sensitivity Test (EFAST) as well as the Sobol method which are well introduced and explained in Saltelli et al. (2008). Saltelli et al. (2010) provides more information about best practices in the calculation of the first and total sensitivity indices. In the next section, the sampling approach of Saltelli et al. (2008) will be introduced as it corresponds to the implementation of the pystran Python Package 4.

5.6.2 Sobol approach for deriving S_j and S_{T_j}

Following the general method described in Saltelli et al. (2008) the calculation of the first and total order indices can be accelerated by the following approach, which uses the quasi-random sampling approach explained in section 3.5. First, perform following parameter sampling and simulations:

- $(N, 2k)$ matrix of random numbers $x_{\mathrm{j}}^{\mathrm{i}}$ using sequences of quasi-random numbers (Sobol, 1967) is generated and divided in two equal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ each containing half of the sample, respectively represented by Equation 5.9 and Equation 5.10.

$$\boldsymbol{A} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_{\mathrm{j}}^{(1)} & \dots & x_{\mathrm{k}}^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_{\mathrm{j}}^{(2)} & \dots & x_{\mathrm{k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_1^{(\mathrm{N}-1)} & x_2^{(\mathrm{N}-1)} & \dots & x_{\mathrm{j}}^{(\mathrm{N}-1)} & \dots & x_{\mathrm{k}}^{(\mathrm{N}-1)} \\ x_1^{(\mathrm{N})} & x_2^{(\mathrm{N})} & \dots & x_{\mathrm{j}}^{(\mathrm{N})} & \dots & x_{\mathrm{k}}^{(\mathrm{N})} \end{pmatrix} \tag{5.9}$$

$$\boldsymbol{B} = \begin{pmatrix} x_{\mathrm{k}+1}^{(1)} & x_{\mathrm{k}+2}^{(1)} & \dots & x_{\mathrm{k}+\mathrm{j}}^{(1)} & \dots & x_{2\mathrm{k}}^{(1)} \\ x_{\mathrm{k}+1}^{(2)} & x_{\mathrm{k}+2}^{(2)} & \dots & x_{\mathrm{k}+\mathrm{j}}^{(2)} & \dots & x_{2\mathrm{k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{\mathrm{k}+1}^{(\mathrm{N}-1)} & x_{\mathrm{k}+2}^{(\mathrm{N}-1)} & \dots & x_{\mathrm{k}+\mathrm{j}}^{(\mathrm{N}-1)} & \dots & x_{2\mathrm{k}}^{(\mathrm{N}-1)} \\ x_{\mathrm{k}+1}^{(\mathrm{N})} & x_{\mathrm{k}+2}^{(\mathrm{N})} & \dots & x_{\mathrm{k}+\mathrm{j}}^{(\mathrm{N})} & \dots & x_{2\mathrm{k}}^{(\mathrm{N})} \end{pmatrix} \tag{5.10}$$

- Define a new matrix $\boldsymbol{C}_{\mathrm{j}}$, constructed by all columns of $\boldsymbol{B}$ except the jth column, which is taken from $\boldsymbol{A}$, as follows:

$$\boldsymbol{C}_{\mathrm{j}} = \begin{pmatrix} x_{\mathrm{k+1}}^{(1)} & x_{\mathrm{k+2}}^{(1)} & \dots & x_{\mathrm{j}}^{(1)} & \dots & x_{\mathrm{2k}}^{(1)} \\ x_{\mathrm{k+1}}^{(2)} & x_{\mathrm{k+2}}^{(2)} & \dots & x_{\mathrm{j}}^{(2)} & \dots & x_{\mathrm{2k}}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{\mathrm{k+1}}^{(\mathrm{N-1})} & x_{\mathrm{k+2}}^{(\mathrm{N-1})} & \dots & x_{\mathrm{j}}^{(\mathrm{N-1})} & \dots & x_{\mathrm{2k}}^{(\mathrm{N-1})} \\ x_{\mathrm{k+1}}^{(\mathrm{N})} & x_{\mathrm{k+2}}^{(\mathrm{N})} & \dots & x_{\mathrm{j}}^{(\mathrm{N})} & \dots & x_{\mathrm{2k}}^{(\mathrm{N})} \end{pmatrix} \tag{5.11}$$

- Based on the resulting simulations performed by all factor combinations in the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}_{\mathrm{j}}$ (every row defines a parameter set to run the model with), three $N \times 1$ vectors are obtained containing the resulting variables of interest. These matrices are defined as $\boldsymbol{y}_{\mathrm{A}}$, $\boldsymbol{y}_{\mathrm{B}}$ and $\boldsymbol{y}_{\boldsymbol{C_j}}$.

Based on the resulting set of vectors, both the first- and total-effect indices can be calculated with a total cost of $N(k+2)$ simulations, which is considerably lower than the N^2 simulations when using the brute force approach. According to Saltelli et al. (2008), the recommended method for estimating the first order sensitivity index S_j is:

$$S_{\mathrm{j}} = \frac{V(E[Y|X_{\mathrm{j}}])}{V(Y)} = \frac{\boldsymbol{y}_{\mathrm{A}} \cdot \boldsymbol{y}_{\boldsymbol{C_j}} - f_0^2}{\boldsymbol{y}_{\mathrm{A}} \cdot \boldsymbol{y}_{\mathrm{A}} - f_0^2} \tag{5.12}$$

with the symbol $(\cdot)$ defining the scalar product and where f_0^2 is defined as follows:

$$f_0^2 = \left(\frac{1}{N} \sum_{i=1}^{N} y_{\mathrm{A}}(i) \right)^2 \tag{5.13}$$

Similarly, the total order sensitivity index S_{T_j} is estimated by:

$$S_{T_j} = 1 - \frac{V(E[Y|\boldsymbol{X}_{\boldsymbol{\sim j}}])}{V(Y)} = 1 - \frac{\boldsymbol{y}_{\mathrm{B}} \cdot \boldsymbol{y}_{\boldsymbol{C_j}} - f_0^2}{\boldsymbol{y}_{\mathrm{A}} \cdot \boldsymbol{y}_{\mathrm{A}} - f_0^2} \tag{5.14}$$

A more in depth discussion about these and other estimators for the first- and total order sensitivity indices is provided by Saltelli et al. (2008) and Saltelli et al. (2010). The practical functionalities for checking the convergence of the indices and visualisation are further described in the pystran Python package 4 documentation.

5.7 Regional Sensitivity Analysis

The Regional Sensitivity Analysis (RSA) is also known as generalized sensitivity analysis or Hornberger-Spear-Young method (Hornberger and Spear, 1981; Spear

and Hornberger, 1980) and is directly related to the Monte Carlo Filtering (MCF) approach (Reichert and Omlin, 1997). Terminology is diverse in this matter, but all of the descriptions basically refer to the same protocol to estimate sensitivity based on any set of simulations resulting from a random sampling procedure (cfr. section 3.5) by dividing the simulations into different groups based on a chosen performance metric. Hence, it is a perfect tool for exploration of the factor space additional to the visualisation of the response surface (Hornberger and Spear, 1981). The method is mainly used in literature for deriving insight into the parameter space (a subset of the possible input factors) towards a performance metric. However, it could as well be used to screen the effect towards any aggregation metric by any input factor, still considering a preferred behaviour to split the factor sets in groups.

Figure 5.5 provides a schematic representation of the individual steps of the RSA method. Similar to all the previous methods, sampling of the input factor space is the starting point of the analysis and for each sampled parameter combination, a simulation needs to be performed. This is illustrated in Figure 5.5a (with the sampling for the input factor θ_j in the small axis). Due to the typical usage of performance metrics, the observations are added as well. The metric values for the simulations are represented by dots in Figure 5.5b. The simulations are divided into a group called *behavioural* (grey dots) and a group *non-behavioural* (black dots) by putting a threshold (horizontal line) on the chosen (performance) metric $V(\hat{y})$. In Figure 5.5b, the the marginal representation for factor θ_1 of the k-dimensional space is shown.

To derive information about the influence of the factors, the empirical cumulative distribution functions are calculated for both groups of factors, represented in Figure 5.5c. A higher number of behavioural parameter values for a range of the input factor will invoke a steeper section in the corresponding range of the CDF. By interpreting the distance between both empirical CDFs, an assessment of the influence of the factor θ_1 can be made. Hence, the distance can be interpreted as a sensitivity index S_1 and a similar figure can be made for every input factor. Figure 5.5d visualises an alternative version with a larger set of groups and showing a smooth version of the empirical CDFs. More details about these alternative representations are explained below.

In the initial contributions by Spear and Hornberger (1980) and Hornberger and Spear (1981), the parameter sets used for the model runs are split into two groups according to their simulation performance: parameter sets which describe the system behaviour sufficiently (*behavioural* parameter sets) and sets which simulate the system insufficiently (*non-behavioural* parameter sets). The frequencies of

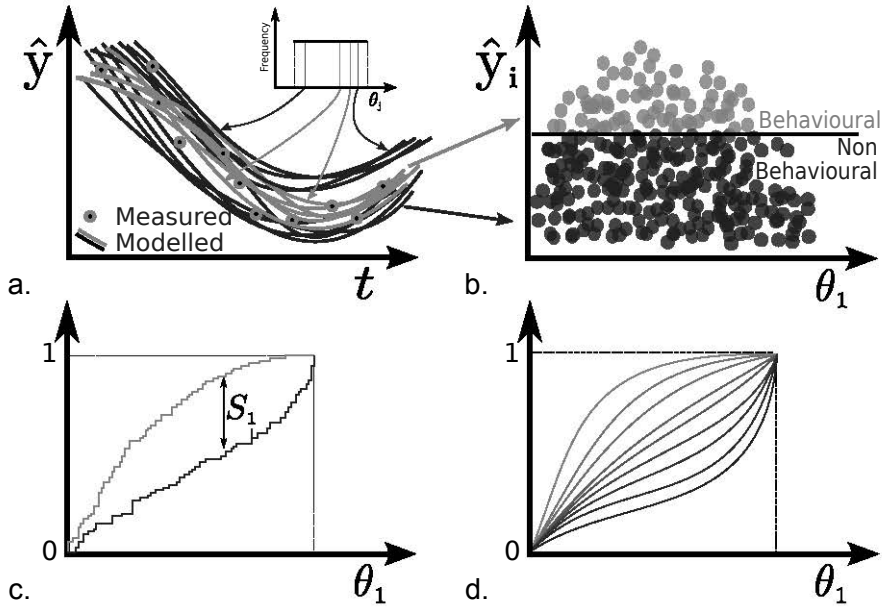


Figure 5.5: Visual illustration of the RSA methodology, which provides mainly a graphical tool to explore parameter sensitivity, by exploring a set of simulations created by a random sampling of the parameter space (**a**). The simulations are divided in two groups based on a decided level of performance (**b**). Graphical representation can be done by comparing the empirical CDFs of both groups (**c**) or by focusing on the empirical CDFs of 10 different subgroups according to their performance (**d**). In Freer et al. (1996), these ten groups are all coming from the *behavioural* group, whereas in the version of Wagener and Kollat (2007) the entire range of the performance is taken into account.

occurrence of the parameter values are accumulated for both groups of parameter sets separately. In other words, the empirical marginal CDFs of both groups are plotted and compared (Figure 5.5c). The distance between the two empirical CDFs can be used as a sensitivity metric. If the two clearly differ, the parameter will be considered influential towards the response variable. Furthermore, the significance of the separation can be estimated using statistical tests such as the Kolmogorov Smirnov (KS) two-sample test and the parameter sensitivity can be ranked using the actual values of the KS measure (Spear and Hornberger, 1980; Saltelli et al., 2008). However, the lack of separation between the CDFs is only a necessary, and not a sufficient condition for non-influence of the parameter. The lack of influence

can also be caused by strong interactions with other parameters (Wagener and Kollat, 2007; Saltelli et al., 2008).

Freer et al. (1996) adapted the initial approach by dividing the group of behavioural parameter sets into 10 equally sized groups based on a sorted model performance metric and comparing the empirical CDFs of these ten sampled subranges. To interpret the qualitative sensitivity of the parameter to a specific performance measure, the degree of dispersion of the ten CDFs represents the influence of the parameter. Wagener and Kollat (2007) used the same idea, but divided the whole range of derived performance metrics into 10 bins and plotted the empirical CDFs with a changing color scheme. These alternatives are represented by Figure 5.5d.

The method mainly provides a graphical support for model structure evaluation. Similar to the regression based SA methodology of section 5.5, it comes free of additional simulation work when a set of simulations is already available, which makes it a helpful additional set of functionalities to have available in the exploration and diagnosis toolset available to a modeller. Furthermore, it provides the basis of the approaches discussed in the next sections.

5.8 DYNamic Identifiability Analysis (DYNIA)

5.8.1 Background of DYNIA

The DYNamic Identifiability Analysis (DYNIA) approach (Wagener et al., 2003) is essentially a dynamic extension of the RSA approach described in the previous section. It can be regarded as the iterative execution of an RSA, where for each iteration the aggregated metric used is a performance metric applied on a small time window. The approach improves the amount of information that is obtained through the use of a moving window (focus on a small subset of the simulation period) instead of an aggregation over the entire simulation.

Similar to the previously described approaches, it uses a large set of simulations based on the sampling of the input factor space, as represented by Figure 5.6a. A major difference with the previous methods is the usage of a pre-defined time-window on which the (performance) metric is aggregated and the split of the input factor sets is done for each individual time-slice, i.e. a moving window before and after each time step. These time slices are visualised in Figure 5.6a and the window around time steps t_c, t_n and t_r are marked in grey.

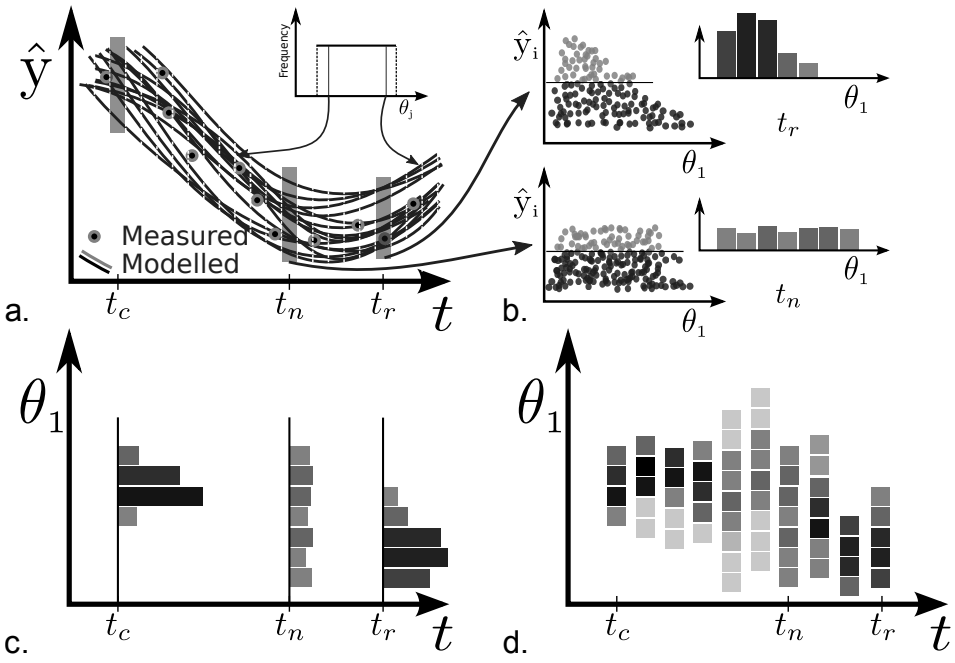


Figure 5.6: Visual summary of the DYNIA approach, starting again from a set of simulations by sampling the input factor space (**a**). For the evaluation, a moving window is used and the filtering principle of the RSA approach is applied on each subset, represented by the time step central of the window (e.g. t_n and t_r) (**b**). The empirical PDF corresponding to each time window is translated into a color intensity corresponding to the density (**c**) after which these are combined in a 2D plot in function of time (**d**).

Analogous to the RSA approach, DYNIA extracts the empirical PDF of the best performing input factor sets. Any performance metric (for which the calculated results can be ranked) can be used to split the *behavioural* from the *non-behavioural* factor values. The main difference with RSA is the explicit usage of a time window. Within each time window, only the best performing factor sets according to a chosen performance metric (e.g. the top 10%) are selected and the empirical PDF is computed based on the metric values. Figure 5.6b illustrates this for the time windows around time steps t_r and t_n. Whereas the RSA is based on the frequency of behavioural factors, the DYNIA approach uses the normalised value of the metric as a weight to derive the marginal PDF.

Looking at the example of Figure 5.6b, the optimal (for the example maximal) values of factor θ_1 are concentrated in a specific region of the parameter space on time step t_r. On the other hand, the optimal values for the window around

time step t_n are not well-defined. This behaviour is reflected in the empirical PDFs shown in Figure 5.6b. Factors that are highly influential for the current time window will be conditioned by the performance metric and deviate from the initial assumed input factor distribution. Darker grey values correspond to higher densities and a steeper gradient of the empirical CDF. The latter is considered as an indicator of the identifiability of the input factor (i.e. parameter) (Wagener et al., 2003).

The results are visualised in a 2D plot of factor values in function of time, where the probability density of the factor is represented by a grey scale, in which a darker grey represents higher identifiability of the input factor for the given time window. Figure 5.6c focuses on the empirical PDF for the time steps t_c, t_n and t_r. The color code is applied to the different bins, with darker shades used for higher densities, as was the outcome of Figure 5.6b. Hence, the input factor θ_1 is more influential (and easier to identify) in the period around time step t_c as it is in the period around t_n. The final representation in Figure 5.6d is based on the color code that represents the factor density. Again, the time steps t_c, t_n and t_r are shown, using the colors of Figure 5.6c. After application of the selection on a time window around each of the time steps and adding these bins to the figure, the resulting graph can be interpreted (see next section 5.8.2).

Furthermore, the 5% and 95% confidence limits of the input factor density function can be calculated and the range is a measure for the ability of the data to discriminate the factor values. Wagener et al. (2003) expressed this in an Information Content (IC) measure as follows:

$$IC_{\mathrm{j}}(t) = 1 - \frac{p_{5\%} - p_{95\%}}{\Delta\theta_{\mathrm{j}}} \tag{5.15}$$

with $p_{5\%}$ and $p_{95\%}$ respectively the lower and upper confidence interval of the obtained marginal input factor distribution and $\Delta\theta_{\mathrm{j}}$ the initial input factor range sampled from. The information criterion ranges between 0 and 1, with high values indicating a small confidence interval expressing high identifiability.

As mentioned, the analysis aggregates the simulations within a specified time window. Hence, for every time step and with a time window of n time steps, the absolute values of scores of the individual time steps between $t - n$ and $t + n$ are aggregated (e.g. summed). The selected time window of the different input factors does not only depend on the influential period of the factor (response time), but also on the quality of the data (Wagener et al., 2003). When the window size is too narrow, the influence of data errors could become too influential, whereas too wide window sizes can result in aggregation of different periods of information

(Wagener et al., 2003). As a rule of thumb, the window size should correspond with the dynamics of the described process. For example, input factors describing fast kinetic reactions will require a smaller window than factors used to describe slow groundwater dynamics. Depending on the window size, the time steps at the beginning and the end of the time series that are distorted need to be excluded from the interpretation (Wagener et al., 2003).

5.8.2 Interpretation of DYNIA

The DYNIA approach originates from the idea that it is needed to evaluate the model during different response modes (Wagener et al., 2001a). Response modes are for example periods of high or low concentration, periods of high or low flow, seasonal periods... However, these periods of interest are not always clearly defined. As such, the DYNIA can be used to identify these different response periods by screening the entire simulation period by a moving window (Wagener et al., 2003, 2004).

A schematic representation of the DYNIA approach plots can be interpreted is shown in Figure 5.7. It provides the representation of the factor influence (behavioural/optimal parameters) in time for the factors θ_1 and θ_2 in combination with the response variable $\hat{y}_i$. The shade of gray represents the density of the factor distribution. When the information content is high, the factor distributions are conditioned within a narrow range with high density (dark grey color). During these periods, the influence of the factor is high and the identification of the factor is potentially possible. In the simplified example of Figure 5.7, these periods seem to align with high values of the variable $\hat{y}_i$, so both factors are important to correctly simulate the behaviour during high values of the variable (e.g. concentrations, flow...).

The light grey areas indicate that equally good parameter values are widely distributed over the feasible input factor range, which corresponds to low sensitivity of the factor and conditions where it will be very difficult to identify unique factor values.

Hence, a first important learning element provided by DYNIA is the identification of periods of high information content for each input factor, taking into account the global input factor range. Hence, comparable to the output of a local SA, sensitive periods can be identified, which supports the process of model calibration and the selection of proper performance metrics, i.e. aggregation towards a set of

periods supporting the model evaluation (cfr. noncommensurable metrics defined by Gupta et al. (1998)).

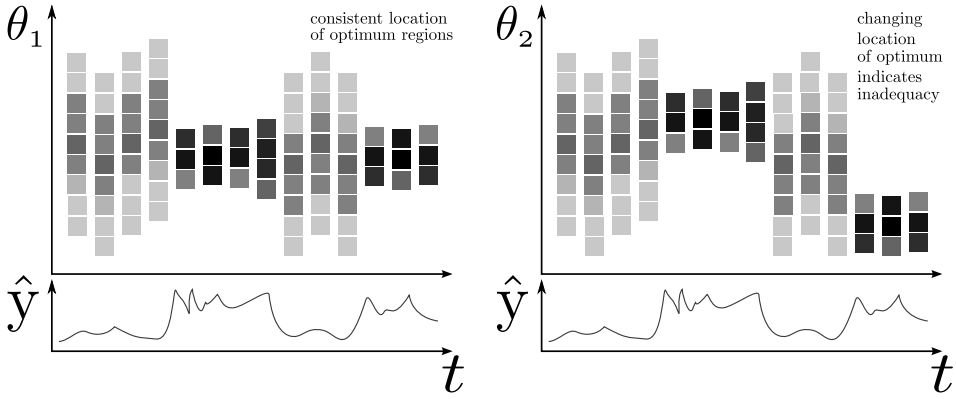


Figure 5.7: Illustration of an idealised outcome of the DYNIA approach, showing the output of 2 factor distributions in function of time for the same period. Darker shades of gray correspond to higher densities. Apparently, both factors are conditioned the most during periods of high values of the variable (sensitive). However, whereas input factor θ_1 is consistently converging to similar optimal values in time, factor θ_2 has other interactions, leading to different regions for optimal parameter values which need further investigation. (Figure is adapted version of the Fig. 4.7 in Wagener et al. (2004))

For input factor θ_1, the region of optimal factor values during these periods is consistent throughout time. In other words, during periods of high values for the variable, the factor θ_1 (part of a model component) has a major influence in a consistent manner. By linking this to the model structure representation, the function of that factor is well identifiable, in line with a parsimonious representation that is looked for. In other words, identifiability can also be interpreted as the property that each factor does have its specific function within the entire model structure and this function can be identified as such. Functions that cannot be identified by the available data should not be included in the model structure.

In the case of factor θ_2, the region of optimal factor values changes in function of time, suggesting interaction with other components during these periods which need to be investigated. In some cases, this can be explained by an interaction with a factor that is part of the same model component, in other applications it is the result of a higher order interaction. Hence, DYNIA indicates a potential inadequacy (interaction effect) of the model that needs to be taken into account

during model calibration or it indicates those model structural elements with room for improvement.

These learning properties make the application of the DYNIA approach well-suited to diagnose model structures, since it relates the deficiencies of the model structure to the adaptation of the input factor values to compensate for this deficiency. This concept will be the central concept in the application of DYNIA in chapter 10, where it is used in combination with other pystran elements.

5.9 Generalised likelihood Uncertainty Estimation (GLUE)

5.9.1 GLUE as model evaluation methodology

The GLUE method (Beven and Binley, 1992; Beven and Freer, 2001) is extending the lack of identifiability of a parameter set of a model structure to the principle of equifinality, which states that multiple input factor combinations of different model structures can give similar (good) model results. The methodology basically selects *behavioural* simulations similar to RSA based on any kind of performance metric and uses the output of these simulations to assess the model output variability (uncertainty).

GLUE is a methodology developed to estimate uncertainty of a model output (Beven and Binley, 1992). However, the applicability of the GLUE method to estimate the prediction uncertainty of a model is prone to debate in literature (Mantovan et al., 2007; Stedinger et al., 2008; Li et al., 2010; Vrugt et al., 2008b; Beven, 2008a; Vrugt et al., 2009). The discussion between formal and informal likelihood functions to estimate the prediction uncertainty (Vrugt et al., 2008b, 2009; Beven, 2008a), is directly linked to the applicability of the GLUE approach. From a metric oriented approach, these are alternative descriptions to quantify model performance. Whereas formal likelihood functions enable also the application of ML and Bayesian methods, informal likelihood functions cannot be used in such rigid theoretical frameworks and the validity of the outcome to estimate the uncertainty when using these informal likelihood functions is questioned (Vrugt et al., 2008b, 2009). Uncertainty analysis refers to some form of quantification, i.e. an estimate of the uncertainty of the model output, which is for GLUE a direct effect of the subjective decision about the threshold for sufficiency (Mantovan and Todini, 2006).

The GLUE method provides an intuitive approach in the evaluation of the effect uncertain inputs have on the variability of the output, conditioned by available data. The fact that the method can be rephrased within a Bayesian context (Sadegh and Vrugt, 2013), a possibilistic context (Jacquin and Shamseldin, 2007) or an approximate Bayesian computation (ABC) context (Nott et al., 2012) illustrates the generic aspect of the idea. Moreover, the fact that the method can be applied to basically any performance metric (or combinations of them, section 3.4) emphasizes its value within a model learning process (Beven and Binley, 1992, 2014). Moreover, it only requires a set of simulations originating from randomly taken samples and is easy to implement. Finally, the method does not make any distinction between realisations coming from different model structures, making it applicable to both input factors (mostly parameters) as well as a set of model structures. As such, it perfectly fits within the diagnostic approach discussed in this dissertation.

Presenting the GLUE approach along with other methods for SA as done here, could be questioned. However, within the context of this dissertation, uncertainty estimation (or error propagation) is not the main objective and GLUE is still useful to gain insight in the model behaviour, similar to the other methods in this chapter.

It should be noted that, in contrast to the presented SA methods, it does not provide information about the individual effect of each input factor, which does not make it a sensitivity analysis according to the definition of Saltelli et al. (2004):

Definition 5.1. *Sensitivity analysis is the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input.*

However, GLUE still provides a straightforward way of propagating the empirical PDF of *behavioural* input factors through a model structure and evaluate the effect of the choices made (performance metric, threshold used, parameter prior distributions...) on the variability of the output. In other words, the GLUE method is able to assess the effect of input factor variability towards the output variability, conditioned by the available observations. Hence, it does have a contribution to determine the influence of model parameters.

When used in the scope of uncertainty analysis, these assumptions should be chosen carefully. The applicability of the GLUE method in the sense of model evaluation is less restricted as it is to apply the method for uncertainty estimation. When testing with different threshold values and performance metrics, the derived uncertainty estimates do still have value in a comparative context, by linking decisions

and uncertain elements to the resulting contribution on the output variability. In this sense, the method is useful to gain insight in the model behaviour without calling it an uncertainty analysis per se.

In contrast to a formal Bayesian approach where the user decides about a specific error function to work with, the flexibility of exploring any performance metric is a major advantage in model structure exploration and diagnosis. In this respect, the further integration with the recently proposed approximate Bayesian computation method is very promising (Sadegh and Vrugt, 2013, 2014). It integrates the flexibility of working with basically any performance metric with the rigorous theoretical framework of Bayesian statistics.

In the context of this dissertation (model learning, testing with different settings), the GLUE method is used in the following sense: provide insight in model behaviour and guidance in the model learning process by model rejection. However, the derived uncertainty bounds should not be interpreted as an estimate of the model prediction uncertainty.

5.9.2 The GLUE approach explained

The major steps when performing a GLUE approach are explained in this section and visualised in Figure 5.8. Similar to the other sampling based methods, the selected input factors to consider as uncertain inputs are randomly sampled with any sampling strategy and from an assumed distribution, as represented by Figure 5.8a. For each of the sampled input factor sets, the selected performance metric needs to be calculated. Each dot in Figure 5.8b represents the resulting metric associated with a simulation. This scatter plot of the performance metric in function of the parameter value is regularly referred to as a *dotty plot* (Beven, 2006).

Subsequently, the user needs to decide about a rejection threshold (or thresholds) to identify non-behavioural model outputs, as visualised in Figure 5.8b. Ideally, the rejection criterion should be chosen before starting the simulations based on the possible observation errors (Pappenberger and Beven, 2006), but in practice the definition of this criterion is mainly a learning process during the analysis itself. When defining the limits of acceptability based on the observation uncertainty (section 3.4.3), the threshold is indirectly defined (Blazkova and Beven, 2009), but relaxation is still considered (Liu et al., 2009b).

The parameter sets with insufficient behaviour (performance values below the agreed threshold) are considered *non-behavioural* and excluded from the subse-

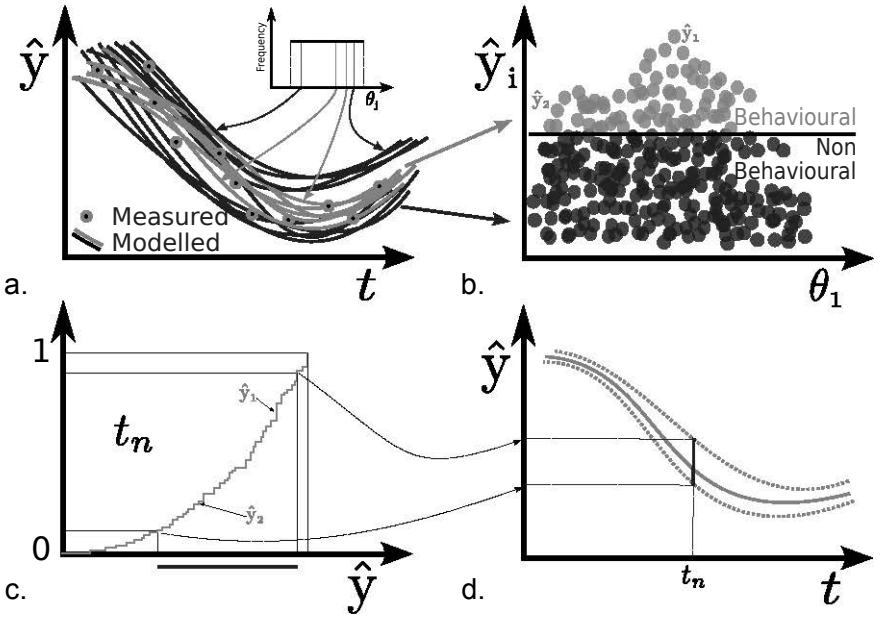


Figure 5.8: Visual illustration of the GLUE approach, starting again from a set of simulations by sampling the input space (**a**). Similar to the RSA approach, the simulations are divided into two groups and only the *behavioural* simulations are used to derive the uncertainty intervals (**b**). For each time step, the empirical CDF is constructed with the normalised perfomance metrics as weights (**c**). By taking the preferred quantiles for each time step, the uncertainty interval can be constructed (**d**).

quent analysis by attributing them a zero weight value in the consecutive steps. Applying this threshold is a crucial step in the analysis, since it is directly related to the final prediction uncertainty. In the utopian situation of exactly one single global optimal parameter set, defining the threshold very strict would result in a brute-force optimization scheme, having left only a single parameter set.

Next, the obtained performance metric values of the behavioural model outputs are normalised. To determine model prediction uncertainty, the model outputs are ranked at every time step and the normalised values are used to construct the cumulative distribution for the output variable, by using the normalised values as weight factors in the empirical CDF. The latter is illustrated in Figure 5.8c for a single time step t_n. The contribution of simulation $\hat{y}_1$ and $\hat{y}_2$ is annotated on both Figure 5.5b and Figure 5.5c. Simulation $\hat{y}_2$ has a metric value just above the threshold value, whereas simulation $\hat{y}_1$ has a larger metric value (higher per-

formance). This results in a larger contribution of simulation $\hat{y}_1$ to the CDF (simulation $\hat{y}_1$ is more likely) in Figure 5.5c.

Prediction uncertainty is subsequently determined by selecting the appropriate percentiles (e.g. 5% and 95%) from the empirical CDF at every individual time step to construct the uncertainty bounds (Figure 5.8d). The term likelihood is not used here, since likelihoods are just one option of the possible performance metrics applied (Romanowicz and Beven, 2006).

5.9.3 Monte Carlo propagation

Another approach for propagating the parameter variability towards the model output is referred as Monte Carlo uncertainty propagation (Saltelli et al., 2008). This is a propagation of the assumed factor distributions by means of a MC approach, resulting in the empirical PDF and CDF of the output variable. So, it is actually a reduced version of the GLUE approach, leaving out the conditioning step of the priori assumed factor distributions by the observations (the factor distributions are assumed to be known).

Similar to the GLUE methodology, the MC propagation approach is a method for uncertainty estimation in the first place. When the probability of the uncertain inputs is known, it provides a straightforward method to derive output uncertainty estimates. In the ultimate version of a completely known description of all the uncertainty input factors by a 'correctly' defined multivariate probability function, a MC propagation approach results in an uncertainty estimation framework. Montanari and Koutsoyiannis (2012) consider this idea as the blue-print for uncertainty estimation, but - personal opinion - which I consider a utopian version of a probabilistic uncertainty approach.

An often seen approach in the application of direct propagation is the usage of a predefined (uniform) variability around the default parameter value with 5%, 25% and 50%, according to the *expected* variability (Reichert and Vanrolleghem, 2001). It is easy to understand and also illustrated by Benedetti et al. (2008) that the determined prediction uncertainty when sampling from an expert-based parameter space is directly linked to the choice of these parameter ranges.

In the specific case of using arbitrary ranges, the propagation provides a method to evaluate how the output variability changes when the variability in the input parameters is altered. When using arbitrary ranges for each of the input factors without taking into account the interactions, unrealistic parameter combinations will be propagated as well. The usage of a conditioning step as it is provided

by the GLUE approach will inherently take into account interaction effects by which some parameter combinations are excluded as they do not provide a proper representation of the observations (Cierkens et al., 2012). Direct propagation of arbitrary input ranges does not account for this. It does provide an idea about how the variability of the model output is triggered by the assumed parameter variability, comparable to the GLUE method. In other words, it provides a technique to assess the effect of a hypothetical uncertainty, exploring potential impact.

Still, the method will be useful to characterize research questions such as: *What could eventually be the consequence on the output when a model parameter would be in reality deviating in a range within 50% of its current estimated value.* When the eventual effect would be very high, then it provides the modeller useful information about the necessity of estimating the particular parameter well. This is similar to factor prioritization as described in other methods for SA. Moreover, for estimating potential risk for basically any *what if?* scenario, the method can be effectively used. But we should be sceptical when communicating about how uncertain one is about a model prediction based on a direct propagation of 'expert knowledge' (arbitrary defined) uniform and uncorrelated factor uncertainties.

5.10 Flowchart for sensitivity analysis

In the previous sections, different methods available to the modeller were explained. These were implemented in the pystran Python Package 4 and some of them will be applied throughout the next chapters. The similarities are apparent and in many cases a combination of different techniques is feasible, certainly for low-dimensional problems. Some guidance as to why a specific method should be selected in a certain situation is helpful.

Still, the flowchart will not keep the modeller from performing the SA method inappropriately, also referred as Type III errors (Saltelli et al., 2008): The usage of adequate factor (parameter) ranges should always be ensured and the unknown effects of those factors that were not taken into account should be considered. Furthermore, it is important to understand that the sensitivity is always the sensitivity as defined by the model structure and not by the natural system modelled.

5.10.1 Selection of a sensitivity analysis method

A flowchart is introduced (Figure 5.9) which can be useful as a practical guidance for applying SA methods. It is not a community-wide agreed flowchart or best practice, but purely a guidance proposed as a starting point. The building blocks were explained in the preceding sections.

However, by making it publicly accessible[2] and adaptable (CC license), it can be further adapted when new methods are developed or other considerations appear. The styling and idea is directly taken from the version originally created by Andreas Mueller in function of the scikit-learn machine learning package (Pedregosa et al., 2011).

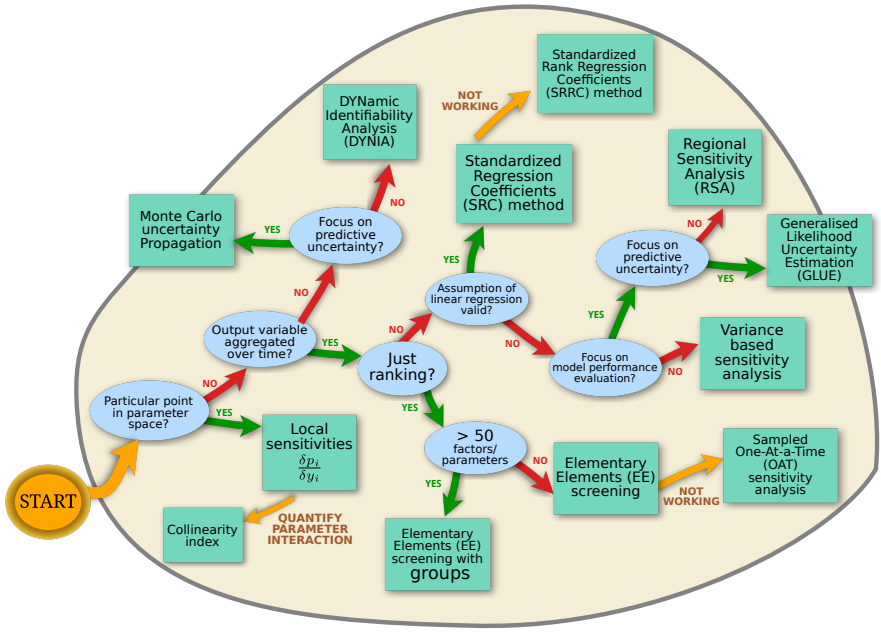


Figure 5.9: Overview and proposal of a decision tree with regard to the methods discussed in section 5.2, providing the basic train of thought to select a specific method of SA. Each of the green boxes corresponds to a well-recognised methodology.

First, the division between local and global methods is made. The former is really easy to apply to any kind of model and can provide already useful information (section 4.2). Collinearity is an extra point of information that comes without extra computational cost.

[2] https://github.com/stijnvanhoey/flowchart_for_sensitivity_analysis

When the entire parameter space is targeted, the first question is if there are already good reasons to decide about a specific aggregation over time (*will aggregating over the entire simulation period be sufficient?*). When not, the DYNIA approach provides an ideal method to get more insight in the parameter behaviour, whereas a pure MC propagation approach will provide information as to how the variability in the output changes as a function of time, considering the assumed parameter distributions. It is noteworthy that other methods can be applied on a time window basis as well and interpreted as such, for example a time-varying variance based method was used to investigate a model structure by Herman et al. (2013b).

When knowledge about useful aggregation functions (which can be performance metrics as well) is already available, it depends on the aim of the SA on what methods to pick. When ranking of the parameters is aimed for, screening methods will suffice. The application of the Morris screening approach using Elementary Effects would be the first pick. For high-dimensional parameter spaces, the application of the Morris method using groups of parameters will be more suitable, however at the cost that parameters within a specific group cannot be ranked. When the Morris method does not provide useful outcomes, a global approach of OAT can be considered as second option. Note that the initial samples of each trajectory can be reused between both methods, but not the whole trajectories.

When (higher order) interaction effects are of interest as well, screening methods will not provide the required answers and other methods need to be considered. In case a linear assumption of the model output in function of the parameters turns out to be reliable, the SRC can be used. Furthermore, ranking the parameter values and simulations can help for non-linear, but monotonic models by comparing the SRRC. However, only qualitative results should be considered when using SRRC.

For many environmental models, non-linearity is common, making a linear regression approach insufficient, leading the modeller to the usage of a variance based approach, which main drawback is the numerous set of simulations needed. When the focus is on checking the effect of parameters on performance metrics, the application of a Regional Sensitivity Approach can provide graphical insight into the response surface to check the effect of individual parameters. When the effect of the conditioning on the model output is aimed for, rather than the parameter characteristics, the GLUE approach is the tool to use.

To emphasise the commonly considered difference between sensitivity and uncertainty analysis, the MC propagation and GLUE approach are put partly outside the area of methods for sensitivity analysis.

5.10.2 Recycling simulations between methods

Instead of the selection of a single method, real benefit would be achieved by integrating these methods in a common workflow, minimizing the total amount of simulations needed to retrieve both the graphical output and the indices of all considered methods. The similar dependence on the sampling of the parameter distributions suggests the potential. For model structures with a limited computational effort and a low-dimensional set of parameters, this is actually rather straightforward (cfr. the assumption in section 3.2), but gets quickly infeasible if the number of dimensions is above 3 or 4.

Hence, the integration brings new methodological and theoretical opportunities. The blending of the Morris screening approach and the variance based approach (Campolongo et al., 2011) and the extraction of a search grid as a byproduct of a sensitivity analysis (Verwaeren et al., 2015) both illustrate that an improved integration is achievable.

Apart from these theoretical efforts, the implementations should be designed to cope with it as well. Hence, the integration of methods and the recycling of simulations among them should be the further development perspective of the pystran and similar environments. The current object-oriented approach of the pystran Python Package 4 is more oriented to the application of a specific method according to the flowchart of Figure 5.9. The modularity of the implementation is mainly supporting the recycling of the code and the coupling with external methods for computing aggregated metrics. As such, it was not fully designed to recycle the simulations amongst different methods. It should be reconsidered in this direction in combination with other packages with similar purposes.

Since the characterisation of the model performance is considered as an iterative process itself (Bennett et al., 2013), we should strive to improve the integration of existing methodologies. The latter is not possible when different algorithms are implemented as standalone executable tools, using various input-output (I/O) file formats and consisting of incompatible (or closed) source codes (Matott et al., 2009). This integration of the pystran functionalities with similar packages is therefore an essential next step. A package such as SALib (Usher et al., 2015), is of specific interest, due to the similar design perspectives as the pystran Python Package 4 and the increasing group of developers.

5.11 Conclusions

This chapter describes the theoretical information of a set of widely used methods for sensitivity analysis which aims to provide future users a sufficient background on the matter to effectively apply these methods. The combination of this detailed description with the online release of the implementation as the pystran Python Package 4, this chapter tries to overcome the typical lack of documentation and transparency regularly encountered Petre and Wilson (2014). The chapter also serves as a reference for the methodologies applied in the other chapters of the dissertation.

At the same time, the necessity of a continuous development to increase the integration of existing and newly developed methods, becomes apparent. This integration should be in terms of the theoretical development as well as in terms of implementation and related documentation of the source code. Alternative implementations are currently available (Ekstrom, 2005; Soetaert and Petzoldt, 2010a; Pianosi et al., 2015; Usher et al., 2015), but a more collective investment of resources should be a next step forward to a community wide library for sensitivity analysis methods. Such an environment could also support the incorporation of newly developed methods such as Mara et al. (2015), while overcoming the overlap between existing frameworks. The proposed flowchart provides a decision tree that gives guidance to novice users in the selection of a method out of the set of methods presented in this work. When adopted by other users, it provides implementation-independent guideline for the application of sensitivity analysis.

PART III

COMPARISON OF HYDROLOGICAL MODEL STRUCTURE ALTERNATIVES

CHAPTER 6

Lumped hydrological model structure VHM

6.1 Introduction

Hydrological models are used to study the potential impacts of future climate change on catchment runoff, to forecast future water levels and as part of integrated modelling studies. These modelling results might be used as a basis for decision making about management of water resources with important consequences for sectors such as agriculture, land planning and water supply. Certainly the adequate prediction of high flows in order to mitigate the risk of flooding and of low flows to assess the impact of droughts is of interest. However, a myriad of hydrological models exists with different levels of complexities and only little information is known about the impact of these differences between the hydrological models on the actual predictions.

Despite the large variety in complexity in hydrological models, an important set of hydrological model structures applied in current research are lumped hydrological models, with a fixed structure based on a certain understanding of the dominant processes in the system (Wagener et al., 2001a). These conceptual models commonly consist of a number of soil water reservoirs and routing routines representing various runoff processes. The Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973), Stanford Watershed Model (Crawford and Linsley, 1966), HYMOD/Probability Distributed Model (PDM) (Moore, 1985), Danish Nedbør Afstrømnings Model (NAM) model (DHI, 2008), IHACRES (Jakeman

et al., 1990) and HBV (Lindström et al., 1997) are some *well-known* lumped model structures regularly seen in literature.

At the same time, these lumped hydrological models are known for their identifiability problems (Beven, 2006, 2008b). Parameter values can hardly be related to physical or measurable properties and observations are essential to inversely identify these parameter values. However, the number of parameters generally limits the practical identifiability of the models, leading to unreliable parameter estimates. As such, the structural property of being an interconnected set of reservoirs with alternative joints and the known issues for identifiability makes this type of hydrological models an ideal case study for model structure identification and evaluation.

The VHM approach (Willems, 2014) used as a starting point in this part of the dissertation is a special case of these lumped hydrological models. It consists of the typical storage and routing blocks like the other models. However, the building process uses another rationale to set up the model structure. It consists of a stepwise approach with a combined model structure and model parameter characterisation. The rationale of the approach is in line with the diagnostic approach of this dissertation, treating model structures as a flexible entity and combining the effort of model structure identification and model calibration (section 2.5). The flexibility makes the VHM modelling approach a good starting case for model structure evaluation. In this part of the dissertation, the flexible model options of VHM will be used to define, implement and evaluate a set of model structural decisions.

The aim of this chapter is to provide the experimental layout on which the following two chapters in this part are based. The study catchment will be shortly introduced, providing information about the natural system studied, i.e. the Nete catchment, together with the available data sets. Next, four different model structure decisions to alter the VHM model will be introduced, resulting in an ensemble of possible model structures. For each of the four model decisions, an assessment of the suitability is aimed for. Finally, the constructed metrics to evaluate the model performance and to assess the model behaviour are explained in more detail. Due to the specific interest of operational water management towards high flows (floods) and low flow (droughts), the constructed metrics are chosen to support these objectives.

6.2 Case study

Study catchment

The Grote Nete catchment, located in the northeast of Belgium with an area of 362 km^2 served as study area as shown in Figure 6.1. It has a temperate climate with an average annual precipitation of 790.3 mm. Rainfall occurs throughout the entire year with more intensive and shorter storms in summer (June till August) and more frequent, generally less intensive, storms in winter (December till February) (Rouhani et al., 2007). The two main tributaries, the Grote Nete and the Grote Laak merge before the Geel-Zammel outlet station. The soils are predominantly composed of sand with around 64% of the area consisting of sandy soils with high hydraulic conductivity (Rouhani et al., 2007). In the southern and valley areas, loamy sand and sandy loam soils are predominant, with minor sections of clay loam and sandy clay loam (Rubarenzya et al., 2007). The topology is rather flat with an average slope of 0.3 % and a maximum one of 5 %, and has a shallow phreatic surface with a water table rising close to the surface in winter. Water resources of the Grote Nete catchment have been profoundly influenced by anthropogenic activities.

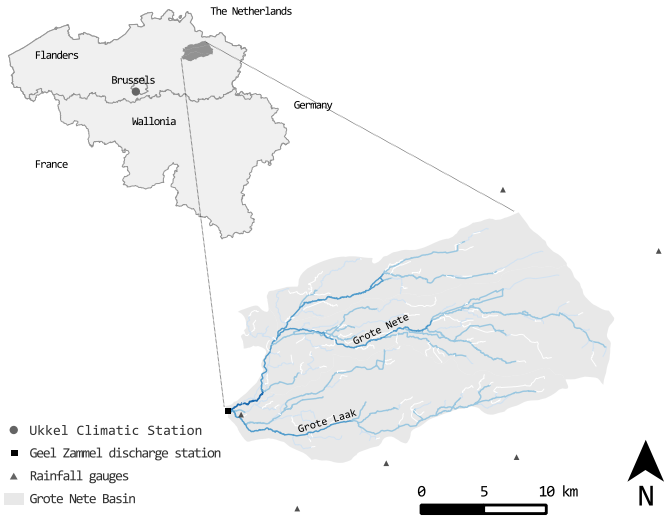


Figure 6.1: Location of the Grote Nete catchment in Belgium and plan view of the river network and gauging stations

Model application observation data preprocessing

Hourly data of rainfall and evapotranspiration are available from different climatology stations in the Flemish area. Lumped hydrological models need a spatially averaged input. Hourly rainfall data was derived by the spatial average of the six neighbouring rainfall gauges shown in Figure 6.1. Evapotranspiration data was measured at the Ukkel climatic station. Daily potential evapotranspiration data were assumed to be representative for the Grote Nete catchment. An empirical relationship to transpose the data to an hourly time step was used (Vansteenkiste et al., 2011) and the resulting time series is shown in Figure 6.3. Hourly flow data of the basin outlet were used to compare the observed and predicted values of the different model structures (Figure 6.2). Based on the availability and quality of the data and in order to span a representative time series, a calibration period from 2002 until 2005 and a validation period from 2006 until 2008 was applied.

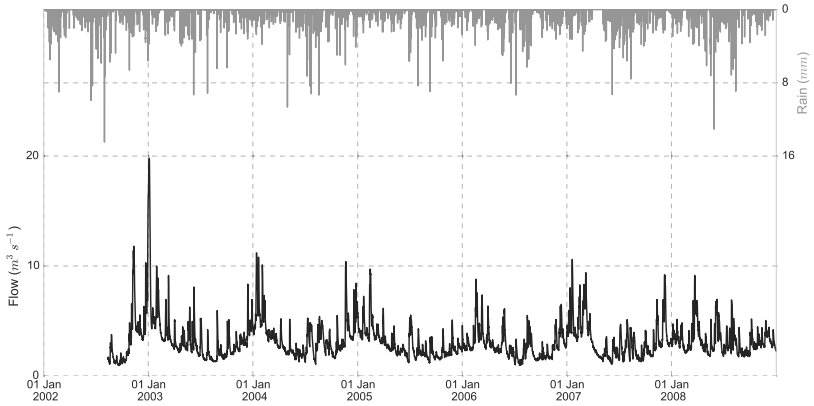


Figure 6.2: Spatial average of rainfall and observed flow at the Geel-Zammel gauging station from 2002 till 2008 as used in the case study

6.3 VHM lumped hydrological model

6.3.1 VHM approach

The VHM approach (Willems, 2014) is a lumped hydrological rainfall-runoff model construction approach. Flexibility is possible, while the possible structural vari-

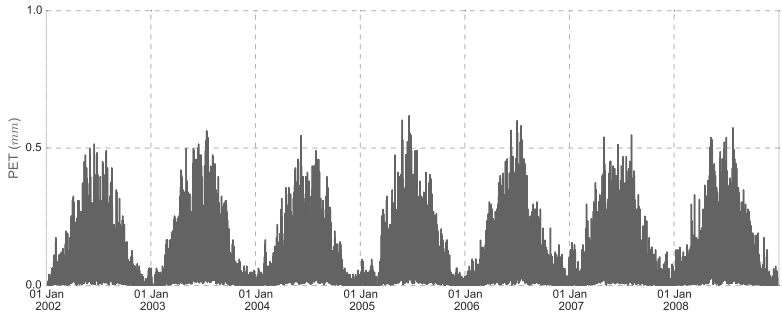


Figure 6.3: Potential evapotranspiration with hourly timestep, derived from the daily time serie available at Ukkel between 2002 and 2008 as used in the case study

ations are kept limited within a specific set of possibilities. The main principle behind all the VHM structure options is the separation of the rainfall into different fractions contributing to the different sub flows by a time-variable distributing valve. The lay-out of the VHM model is shown in Figure 6.4 and a detailed description about the modelling approach can be found in Willems (2014).

The entities of the model are the soil storage defining the dynamics of the soil water storage compartment combined with a number of (linear) reservoirs defining the routing part of the model, comparable with other lumped hydrological models with a soil storage section and a routing section as main entities (Kokkonen and Jakeman, 2001). The balance of the soil storage is given by

$$\frac{\mathrm{du}}{\mathrm{d}t} = \mathrm{p_{t,in}} - \mathrm{q(u)} - \mathrm{e(u)} \tag{6.1}$$

with u (mm) the soil moisture storage, $\mathrm{p_{t,in}}$ ($\mathrm{mm\,s^{-1}}$) the rate of rainfall (intensity), q ($\mathrm{mm\,s^{-1}}$) the runoff rate generation and e ($\mathrm{mm\,s^{-1}}$) the actual evapotranspiration rate. The outgoing fluxes e and q are both function of the soil moisture storage. The transformation from potential evapotranspiration to actual evapotranspiration is assumed to be linearly related to the soil moisture storage for all models in this study, but can be varied for other applications. The runoff rate generation q is split into different sub flows. Flows are calculated based on the attributed fractions from the rainfall with $\mathrm{q(u)} = f(\mathrm{u}) \cdot \mathrm{p_{t,in}}$ resulting in the flow $\mathrm{q_{of}}$ for overland flow, $\mathrm{q_{if}}$ for inter flow and $\mathrm{q_{bf}}$ for base flow. Overland flow is the direct runoff, the inter flow conceptually represents the subsurface flow which influences

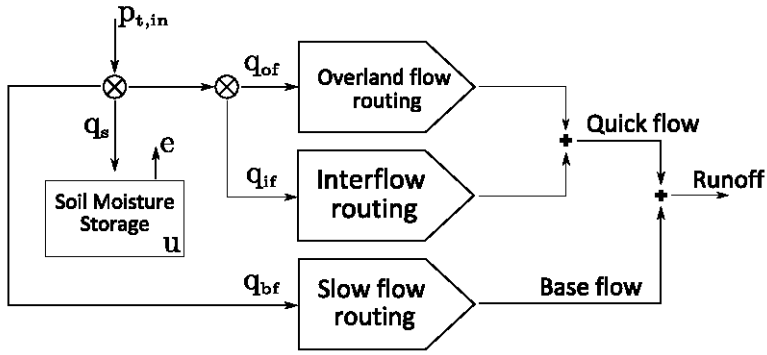


Figure 6.4: VHM model structure, representing the main building blocks of the model and fluxes calculated by the model

the runoff characteristics of the catchment as well and the base flow describes the groundwater contribution of the flow.

The function $f(\mathrm{u})$ computes the fractions and can be adapted in function of the representation and process description. The function descriptions will be introduced in section 6.3.3, for example Equation 6.2 and Equation 6.3. Mass balances are closed at all times by verifying that the sum of the fractions equals 1 at all times.

6.3.2 Implementation of the VHM model structure

Python implementation

The flexible nature of the model structure identification described in Willems (2014) is the basis for the defined model structure alternatives.

The original implementation of the model is not open and is programmed in Excel with Visual Basic. The latter prohibits the access of the code and the handling of the model in connection with other components. As such, the implementation of the flexible approach of the VHM was done in the scripting language Python, to increase the flexibility and extendibility of the model. The model is available as a single Python function, see Python Module 5.

Python Module 5 (VHM_flexible)**.**
Flexible and straightforward application of the VHM modelling approach as proposed by Willems (2014). More information about the implementation is provided together with the code at `https://github.com/stijnvanhoey/flexible_vhm_implementation`

Model output showcase

An introduction of the model is provided for a single model structure output to show the different outputs that the model implementation provides. The discharge of the gauge Geel Zammel in the Nete catchment is used and the modelled discharge is shown together with the observations in Figure 6.5. The parameter set is retrieved from the original workflow described in Willems (2014) and performed by Vansteenkiste et al. (2011). Overestimation of the lower flows during summer months and underestimation of the winter peaks in 2004, are major deficiencies noticed for this calibration.

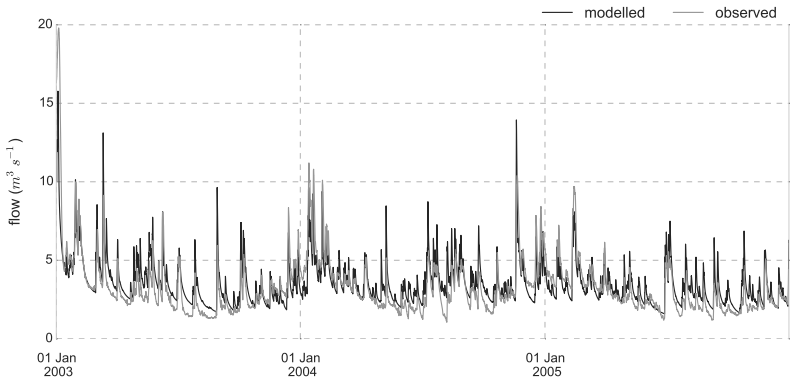


Figure 6.5: VHM modelled flow example in comparison with an observed time series of the catchment

As noticed in Figure 6.4, the model assumes three subflows that contribute to the total flow. Hence, sub flow filtering techniques can be applied to the observed time series to estimate the proportional part of each of these subflows (Willems, 2009). Figure 6.6 compares both the modelled and filtered observed subflows. Except for the underestimated base flow in winter months, the model output captures the dynamics in the different subflows well using the parameter set of Vansteenkiste et al. (2011) (also provided in Table 7.3). Combination of Figure 6.5 and Figure 6.6 learns that the overestimation in the summer months is mainly caused by the

mismatch of the inter- and overland flow. Still, the subflow filtering on itself is also uncertain and dependent on the chosen filter and settings.

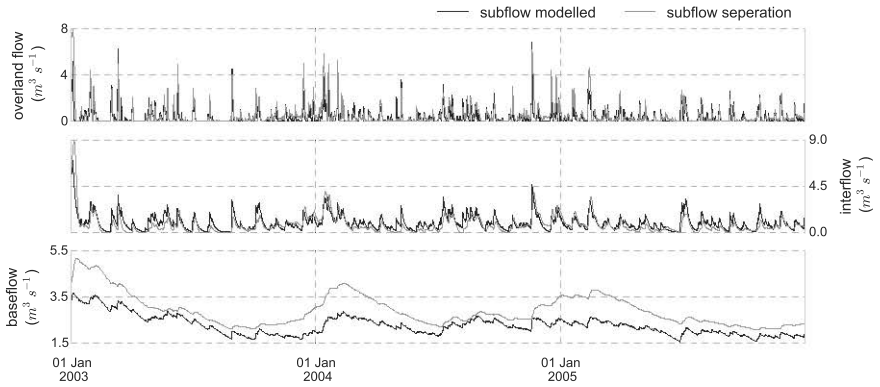


Figure 6.6: VHM subflow modelling example in comparison with a subflow separation result of the flow time series, done with a numerical filter described in Willems (2014)

The VHM approach essentially performs a rainfall fractionation determining the redirection of the incoming rainfall towards the different components. In this sense, the model is different in comparison to other lumped hydrological models (Moore, 1985), since these typically do not directly pass on a fraction of the rainfall towards the base flow component. Still, the fraction of the rainfall contributing to the soil moisture depends on the state of the soil moisture at each time step, which is provided in Figure 6.7.

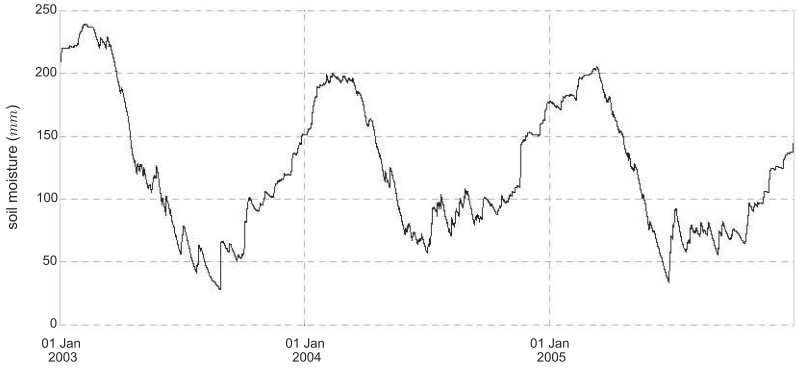


Figure 6.7: VHM output of the soil moisture storage in function of time, representing the moisture state of the Nete catchment.

Figure 6.8 and Figure 6.9 provide both an overview of the fractionation in function of time, respectively with and without taking into account the description of infiltration excess expressed as a function of the antecedent rainfall at each moment. In contrast to saturation excess, which is conceptually represented by the filling of the soil moisture storage, the infiltration excess describes the runoff initiated when the rainfall capacity is larger than the soil infiltration capacity.

As a first check, the sum of the fractions should be unity, which is correct for the entire simulation period as illustrated by the constant value for the sum of the fractions in Figure 6.8. The other fractions in Figure 6.8 are not smooth and shows a spiky behaviour which seems hardly realistic. The latter is due to the conceptualisation of the infiltration excess by using antecedent rainfall in the product for both overland flow and inter flow, leading to these sudden drops in the fractions of overland flow and inter flow. When there is no antecedent rain during the chosen time period, the overland flow and inter flow fractions instantly drop to zero, leading to a sudden decrease. When rain starts again, the fractions immediately increase again as the term in the product is no longer zero (see also Equation 6.5 in the next section). Since the base flow is implemented as a rest fraction of the other fractions, it increases at the same time.

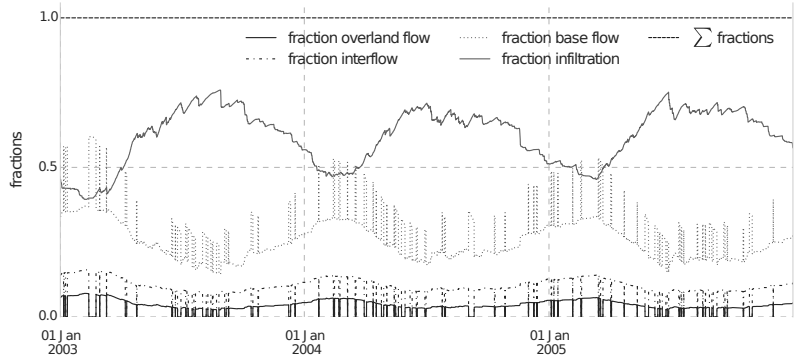


Figure 6.8: VHM output of the different fractions contributing to the subflows, when the antecedent rain concept representing the process of infiltration excess is included in the model structure.

6.3.3 Implemented model component adaptations

Different types of model structural changes can be applied to the VHM model structure. Within the scope of this application, the focus is not to generate an

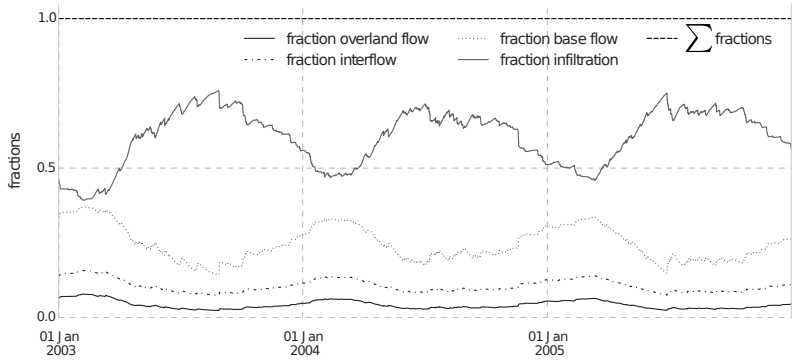


Figure 6.9: VHM output of the different fractions contributing to the subflows, when the antecedent rain concept representing the process of infiltration excess is excluded from the model structure.

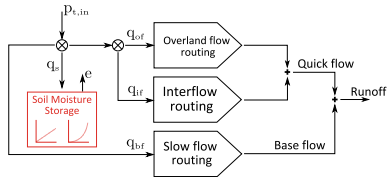
extensive number of model structures, but rather to discriminate between a small number of rival model structures that can be interpreted as equivalent representations of the system. Four types of model adjustment were identified and chosen for the further analysis. Each of the four adjustments are linked to a specific model structure decision and represent each a different type of model structure manipulation. In this section, the four selected adjustments will be discussed in more detail, each linked to a type of structure manipulation.

We define a model component as a conceptual description of a (sub)process of the entire model. This can either be the mathematical description of a specific flux (e.g. percolation, evapotranspiration...) or an entity in the model represented by a mass balance (e.g. upper soil layer).

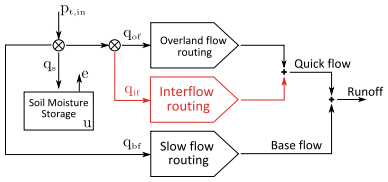
- **Change component mathematical structure**: The mathematical formulation can be altered, by which the relationship between the variables is changed. As such, the parameter values are different, but can retain their referred (physical) representation. This can be combined with an increased number of parameters, but is not necessarily the case.

  Implemented for this case (Figure 6.10a) is the transformation from a linear relationship between the soil storage flux fraction and the soil storage

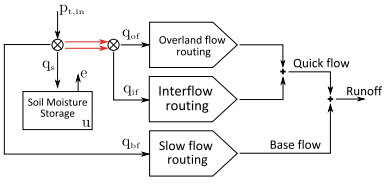
$$f_{u,1}(t) = s_1 - s_2 \frac{\mathrm{u}(t)}{u_{\max}} \tag{6.2}$$



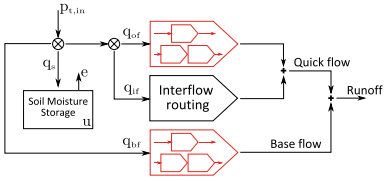
(a) Linearity of the soil moisture storage



(b) Necessity of the inter flow process



(c) Incorporation of the antecedent rain concept to incorporate infiltration excess



(d) Representation of the routing components

Figure 6.10: Application of the different model variation on the VHM model structure, with the structure adaptations marked in color

towards a non-linear equation using an exponential dependency and addition of an extra parameter:

$$f_{u,2}(t) = s_1 - e^{s_2\left(\frac{u(t)}{u_{max}}\right)^{s_3}} \tag{6.3}$$

The concepts behind parameters s_1 and s_2 are similar, with s_1 defining the maximum fraction (dry soil) and s_2 the gradient of the function, defining the minimum fraction (saturated soil). Parameter s_3 defines the curvature of the function in the non-linear case. As such, the higher complexity defining the non-linear relationship defines more degrees of freedom to mimic the 'real' soil storage infiltration function.

- **Delete model component process**: Deleting a specific model component is a direct way of model structure reduction. Nevertheless, it does not always mean that the process is not occurring, but it is assumed to be of minor relevance for the specific purpose. The reverse action, adding a component, is similar, but the reasoning is opposite. The central issue is whether the added component can be parametrically identified. In general, this type of model structure comparison will be the case if model reduction is intended.

  To represent this reduction, leaving out the inter flow component was proposed as a possible model reduction (Figure 6.10b). Since the inter flow

conceptually represents the subsurface flow, it is assumed that the effect of the subsurface flow has no major impact on the runoff characteristics of the catchment.

- **Increase component variable dependency**: A specific process is dependent on certain variables, but the dependency structure can be changed by extending the equation with an extra variable. By doing so, the described process is assumed to be influenced by the added variable whereas it was not in the original structure.

  As an example, the fraction contributing to overland flow and inter flow are described in two distinct ways (Figure 6.10c). In a first variation, the dependency is only based on the characteristics of the current soil water storage.

$$f_{of,1}(t) = o_1 e^{o_2 \frac{\mathrm{u}(t)}{u_{\max}}} \tag{6.4}$$

  As an alternative, $s(t)$, i.e. the antecedent rainfall volume, is introduced as an extra variable to express the dependency of overland flow on the total rainfall budget of the $n_\circ$ previous time steps.

$$f_{of,2}(t) = o_1 e^{o_2 \frac{\mathrm{u}(t)}{u_{\max}}} s(t)^{o_3} \tag{6.5}$$

  The parameter $n_\circ$ defines the number of time steps used to compute the cumulative rainfall for, resulting in $s(t)$. This represents the wetness of the soil surface and can be seen as the addition of an infiltration excess term in the equation, whereas the other part represents the saturation excess process. Parameter o_3 is added as an extra parameter, giving two extra parameters in total for the overland flow. A similar substitution was considered for the inter flow model equations.

- **Extend component parameterisation or variables**: Sometimes the difference between model structure variation and changing parameterisation is not evident and depends on the implementation. As a straightforward example to clarify this, the routing concept in lumped hydrological models is explained in more detail.

  A linear reservoir is used regularly to describe the routing of water and to simulate the retention process of water moving to the outlet. A single reservoir is characterised by the retention parameter K (storage factor) and the equation can be solved analytically. However, this concept is generalised in

the so-called Nash cascade of linear reservoirs, each having the same storage factor. As such, the resulting mathematical structure can be solved analytically and the resulting unit hydrograph is comparable to a two parameter Gamma distribution, defined by a storage factor K and a number of reservoirs n (no longer limited to integer values). As such, the Nash-cascade based routing model structure can be defined as one component, defined by two parameters (K and n). Nevertheless, when solved numerically, the number of reservoirs (n) defines the number of mass-balances in the model component to describe routing. This means that the model is extended with extra state variables when more reservoirs are included.

In this case, we will consider the increase of the number of reservoirs as adding extra variables to the model, since it defines the most general way of model extension. Each reservoir defines a separate mass balance and can be defined by a unique storage factor K_{i} and can be described both in a linear or non-linear way. A Nash cascade of linear reservoirs is just a special case, where all reservoirs are selected to be linear with the same storage factor.

As an example, both the overland flow and base flow were varied between one or two linear reservoirs, each having a separate storage factor, represented in Figure 6.10d.

In summary, based on the former classification of model structures, twenty four (24) rival models were constructed for the study by making the combination of these model structure options.

Figure 6.11 represents the four different conceptual model decisions and how the combination leads to the studied ensemble of model structures: (1) the relationship of the soil storage component can either be defined as linear or exponential; (2) inclusion or exclusion of an inter flow component to represent drain flow and runoff from the vadose zone; (3) the surface runoff submodel was extended with an infiltration excess submodel or not; (4) the configuration of routing reservoirs of the base flow and overland flow routing by extending one or the other with an extra reservoir. For each of these model decisions, the model options are listed in Figure 6.11, together with a label in *italic* to make identification of each individual model structure possible. For example, the model structure *exp ni na no* defines a model structure with an exponential storage compartment, without interflow, without infiltration access and without additional routing reservoirs taken into account. Combination of these options results in the 24 model structures used in the ensemble to evaluate, which is represented by the lines in between the model options.

The model parameters are listed in Table 6.1 with for each parameter a minimum and maximum value acting as the boundaries of the parameter space. Moreover, the model component of which the parameter is part of is given, together with a label for each of the components. The labels of the components will be used later in chapter 8.

All model structures have a single soil moisture storage component and a linear transformation of potential evapotranspiration to actual evapotranspiration. As such, the effect of the four model options can be tested by a comparison of 12 model configurations for option (1) to (3) and for six configurations for option (4), since the latter is split to three options.

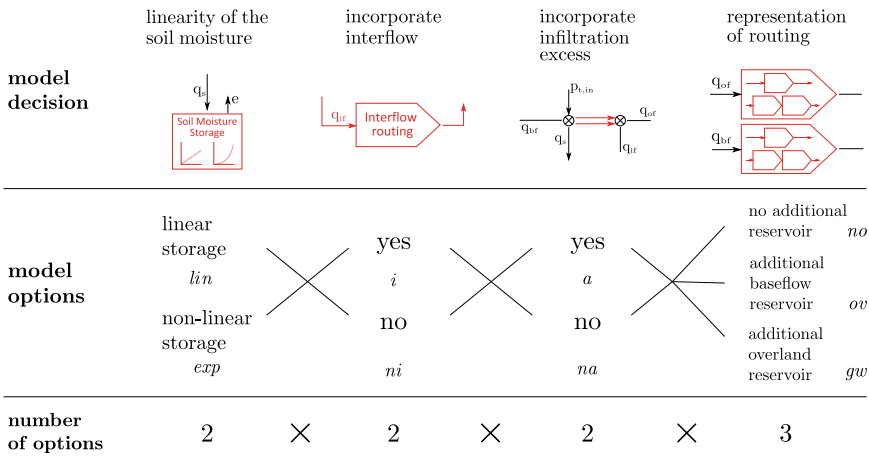


Figure 6.11: Overview of the created ensemble of model structures based on the variation provided by the four selected model structure adaptations to the VHM model. A label is assigned and added in *italic* to each model structure option to enable the identification of each model structure combination by combining the labels. The lines represent the different combinations of model options to construct a model structure. The combination of the different model decisions results in an ensemble of 24 model structures.

6.4 Performance metrics

Chapter 3 discusses the importance of selecting appropriate performance metrics to support the research question. Three different performance metrics were defined to evaluate the model performance for this application, each representing a different model objective. A reliable prediction of high and low flows are of specific interest

Table 6.1: Overview of model parameters, the ranges of variation and the related component. Common parameters are present in all the 24 model structures. The last column provides the label of the component, used in the graphical representation in chapter 8.

Parameter	Minimum	Maximum	Component	Label
$u_{\max}$	200.0	500.0	Storage	S
s_1	1.0	3.0	Storage	S
s_2	0.1	2.0	Storage	S
s_3	0.1	2.5	Storage	S
u_{evap}	90.0	250.0	Evap	E
o_1	-6.0	-3.0	Overland	O
o_2	1.0	6.0	Overland	O
o_3	0.2	2.0	Overland	O
n_{o}	3.0	48.0	Overland	O
K_{o1}	10.0	120.0	Overland	O
K_{o2}	10.0	120.0	Overland	O
i_1	-6.0	-3.0	Inter flow	I
i_2	1.0	6.0	Inter flow	I
i_3	0.2	2.0	Inter flow	I
n_{i}	3.0	48.0	Inter flow	I
K_{i}	90.0	150.0	Inter flow	I
K_{g1}	1500.0	2500.0	Base flow	B
K_{g2}	1500.0	2500.0	Base flow	B

for operational management, as both are linked with potential threads, respectively floods and droughts.

As a first metric, the frequently applied NSE (Nash and Sutcliffe, 1970) was used, mainly as a well known reference. NSE tends to overestimate the deviation between modelled and measured values of high flow peaks. To draw attention towards lower flow values, adapted versions of the Nash-Sutcliffe criterion can be used in order to less heavily penalize large differences (Schaefli and Gupta, 2007). For the analysis presented here, two alternative performance criteria to emphasize respectively low and high flow conditions are designed. For the design, the aim is to have metrics that are not based on the comparison of individual time steps, as a small shift in time in between the observed and modelled time series can result in bad performance, whereas the dynamics could be well captured (Dawson et al., 2007). At the same time, specialisation towards either low or high flows should be supported as well.

The designed metrics are based on the Flow Duration Curve (FDC) derived from the modelled and measured time series. Instead of constructing the metric based on the residuals as a function of time, the evaluation is done by comparing the FDC of both the observed and modelled time series. Residuals are calculated by

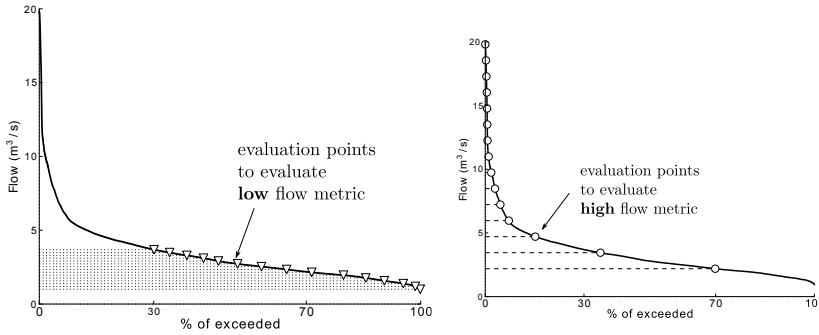
subtracting corresponding values of the observed and modelled FDC, which makes the evaluation time-step independent.

Similar metrics were applied by Westerberg et al. (2011b); Blazkova and Beven (2009); Refsgaard and Knudsen (1996), who compare the FDC of the simulated flow and the observed flow in discrete points along the FDC. In this work, this concept has been further extended to support the specialisation towards either low or high flows. By choosing the evaluation points along the FDC at the lower or higher quantiles of the flow duration curve, emphasis is given respectively towards high (called HighFlow) or low (called LowFlow) flow values. The transformation for both the modelled output and the observed data from the original time series towards the evaluation points can be regarded as an aggregation function to derive the metric. The translation towards a performance metric is done by comparing the resulting evaluation points of both for which a choice can be made of a specific category of metrics as listed in section 3.4.1, which is done here comparable to the definition of the NSE (comparison of residuals with the residuals of the mean as reference model).

Figure 6.12 shows the evaluation points of the flow duration curves for the LowFlow and HighFlow metrics. The choice of the range to spread the evaluation points in both criteria, is made based on the analysis of the hydrograph of the study catchment and should be verified when used for other catchments. For the low flow performance criteria discrete evaluation points (EP) are taken between the 30% and 100% quantiles focusing on the base flow values and for the high flow performance criterion between the 0% and 70% quantiles to emphasize peak flows and the recession after a rain event. By choosing this division, both criteria are interacting for around 70% of the time steps of the total observed time series.

To ensure that the discrete points capture the curve of the flow duration, the evaluation points were evenly distributed between the minimum and maximum measured flow value of the selected quantiles. By doing so more emphasis is going to strongly changing parts of the flow duration curve (evaluation points are closer to each other in Figure 6.12). Using an equidistant distribution of evaluation points between the minimum and maximum quantile values would result in only few evaluation points in the steeper parts of the curve, whereas these are of most interest.

Finally, this set is extended with a combination of NSE on the one hand and the LowFlow and HighFlow on the other hand, both equally weighted and further respectively called NSE-FDClow and NSE-FDChigh, available as well in the set of performance metrics in the pystran Python Package 4:



(a) evaluation points for low flow (range 30-100%)

(b) evaluation points for high flow (range 0-70%)

Figure 6.12: Illustration of the Evaluation Points (EP) used to derive the FDC based performance metrics for respectively low flow (left) and high flow (right). To calculate a metric, the FDCs of both the modelled and observed time series are derived and the residuals are calculated for each of the flow values that corresponds to an evaluation point. When focusing on low flow values (LowFlow), the range is 30-100%, whereas for high flow values (HighFlow), the range is 0-70%.

$$\text{NSE-FDClow} = w_1 \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} + w_2 \frac{\sum_{l=1}^{M}(\hat{EP}_l - EP_l)^2}{\sum_{l=1}^{N}(\bar{EP}_l - EP_l)^2} \tag{6.6}$$

with M the amount of evaluation points (EP) chosen on the FDC in between the 30% and 100% quantiles. The calculation of the NSE-FDChigh only differs in the chosen evaluation points on the FDC. w_1 and w_2 are the weights that can be attributed to each of the terms with a default value of 1. In total, 5 different performance metrics are defined, i.e. NSE, LowFlow, HighFlow, NSE-FDClow and NSE-FDChigh.

To illustrate the alternative focus of the LowFlow, HighFlow and NSE metrics, a scatter diagram of the calculated metric values for a set of simulations resulting from a randomly drawn set of parameters is shown in Figure 6.13. In this figure, the NSE is adapted to make sure lower values are representing a higher performance (1-NSE of only the strict positive values) to make all the evaluation criteria similar in interpretation. Clear correlations of the clouds would have revealed redundancy of the different performance functions. Hence, the drafted LowFlow, HighFlow and the NSE criteria are focusing on different aspects of the hydrograph. The importance of providing performance criteria with strong discriminatory power is addressed by Kavetski et al. (2011) and Gupta et al. (1998). The Transformed

Root Mean Square Error (TRMSE) (Wagener et al., 2009) is also included in the figure. Low flow observations are higher weighted in the TRMSE evaluation, so having a similar focus as the LowFlow metric defined earlier. The similar focus is represented in the graph by the high correlation between the FDC-based LowFlow metric and the TRMSE. A trade-off between low and high flow criteria support the idea that the performance metrics are able to discriminate through their specific characteristics.

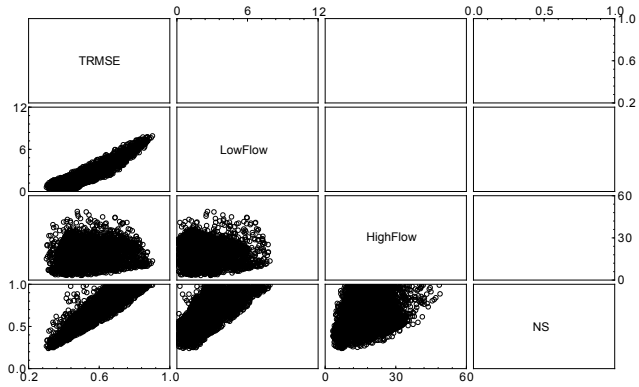


Figure 6.13: Example performance criteria scatter plots showing the correlation and the trade-off between the LowFlow, HighFlow and NSE performance metrics (lower values represent a better fit for all criteria). Each subplot represents the scatter plot between two of the evaluation criteria used (TRMSE is added to check the similarity with the FDC LowFlow criterion) for the simulations.

6.5 Conclusion

The elements introduced in this chapter are the building blocks for the remainder of this part of the dissertation, in which the implemented flexibility of the VHM approach is investigated and evaluated. The chosen model decision to assess are constructed and the performance metrics to support either low flow or high flow as research objective are defined.

The structures that are part of the ensemble resulting from these four model structure decisions are considered to be equivalent representations of the system and will be evaluated in chapter 7 and chapter 8 by using the defined performance metrics.

In chapter 7, the focus is on the performance of the individual model structures that are part of the ensemble and the ability to distinguish the model structure based in terms of model performance. In chapter 8, the shift in the sensitivity of the parameters when introducing model alternatives is used to derive information about the model structure behaviour and the suitability of the individual model decisions.

CHAPTER 7

Ensemble evaluation of lumped hydrological model structures

Redrafted from

Van Hoey, S., Vansteenkiste, T., Pereira, F., Nopens, I., Seuntjens, P., Willems, P., and Mostaert, F. (2012). Effect of climate change on the hydrological regime of navigable water courses in Belgium: Subreport 4 Flexible model structures and ensemble evaluation. Technical report, Waterbouwkundig Laboratorium, Antwerpen, België

7.1 Introduction

In this chapter, it is questioned whether the provided set of model structures within the ensemble created in the previous chapter can be distinguished using the defined set of performance metrics and based on an optimization approach. Furthermore, the lack of parameter identifiability and the dependence of the resulting calibrated parameter set on the decided performance metric is illustrated.

First, the effect of the used performance metric on the resulting calibrated parameter set is tested by calibrating a single model structure towards different performance metrics. Subsequently, the optimization is used to calibrate the 24 different versions of the VHM model structure defined in the previous chapter and compare the performance of the individual structures. The aim is to check if specific model decisions lead to better performance and can be distinguished as such.

7.2 Effect of performance metric on model calibration

To compare the parameter sets resulting from model calibration performed with different performance metrics, the VHM structure configuration used in Vansteenkiste et al. (2011) was selected, i.e. the *exp i a no* (see Figure 6.11 for the component option labels). The structure has a non-linear storage component, both interflow and antecedent rain included, but no extra routing complexity.

The purpose is to find the optimal parameter sets when using the selected performance metrics. Hence, this is an optimization problem for which the SCE-UA optimization method will be used. SCE-UA was introduced in chapter 3 and is available as Python Module 3. It is a global optimization algorithm proven to be effective and efficient in locating the globally optimal model parameters of a hydrologic model (Duan et al., 1992). Using the same model structure as the one used in Vansteenkiste et al. (2011) provides the possibility to compare the resulting parameter sets with the manual calibration results derived in that study. To do so, the same split sample approach and corresponding periods were used as in Vansteenkiste et al. (2011), i.e. the dataset starting from August 2002 until the end of 2005 for calibration, covering a wide range of climatic and hydrological conditions.

The performance metrics provided in chapter 6 are separately used as single criterion to calibrate the model. So, in total, the automated optimization is performed using five different performance metrics (NSE, LowFlow, HighFlow, NSE-FDClow and NSE-FDChigh), resulting in five separate optimization problems.

In Tables 7.1 and 7.2 the performance metrics that were used in Vansteenkiste et al. (2011) to evaluate the hydrological model simulations, are used to compare the results of the different optimizations in terms of general performance as defined by other metrics. Vansteenkiste et al. (2011) expressed the performance by the coefficient of determination (R^2), the mean absolute error (MAE) and the root mean squared error (RMSE) (for which a description was given in section 3.4). The equivalence of the metrics using the NSE and the minor overall performance of the LowFlow and HighFlow metrics is noticed (lower NSE and R^2, higher MAE and RMSE). However, the relative decrease in performance in between the calibration and validation set is less prevalent for the LowFlow and HighFlow metrics, indicating their robustness and the potential exaggeration in the fitting of the other metrics.

The validation is done to evaluate the usefulness and predictive power of the optimized parameter combinations outside the calibration period. A comparison of the observed and modelled discharges over the 3-year calibration period and subsequent 3-year validation period downstream the catchment for respectively NSE, HighFlow and LowFlow is shown in Figure 7.1.

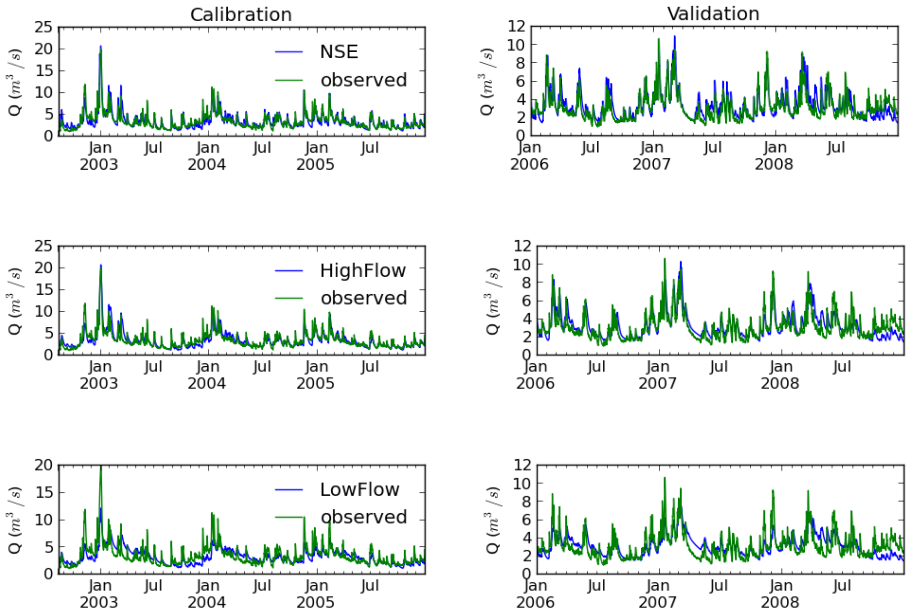


Figure 7.1: Observed (green) and simulated (blue) runoff discharges downstream the catchment based on the automated calibration for both the calibration (left) and validation (right) period. The first line shows the best performing simulation using the NSE performance metric, the second line the best performing simulation using the HighFlow performance metric and the third line using the LowFlow performance metric.

Figure 7.1 demonstrates that the runoff predictions for each of the models depends on the performance metric used during calibration. This is quantified by the performance metrics presented in Tables 7.1 and 7.2. From Figure 7.1 it can be seen that the use of the FDC-derived performance metrics (LowFlow and HighFlow) results in model realisations that are less suitable to grasp the overall dynamics of the streamflow. Especially the LowFlow metric is underestimating the peaks and overestimating the recession periods from January to June, which translates to lower NSE and R^2 values and a higher mean absolute error (MAE).

Table 7.1: Performance values for the calibration period (2003-2005) after calibration with regard to different metrics: Nash-Sutcliffe coefficient (NSE), the regression coefficient (R^2), mean absolute error (MAE) and the root mean squared error (RMSE)

	NSE	LowFlow	HighFlow	NSE-FDClow	NSE-FDChigh
NSE	0.85	0.57	0.75	0.85	0.84
R^2	0.92	0.76	0.88	0.92	0.92
MAE	0.57	0.94	0.7	0.57	0.58
RMSE	0.8	1.35	1.03	0.81	0.82

Table 7.2: Performance values for the validation period (2006-2008) after calibration with regard to different metrics: Nash-Sutcliffe coefficient (NSE), the regression coefficient (R^2), mean absolute error (MAE) and the root mean squared error (RMSE)

	NSE	LowFlow	HighFlow	NSE-FDClow	NSE-FDChigh
NSE	0.7	0.51	0.61	0.69	0.67
R^2	0.88	0.71	0.81	0.88	0.87
MAE	0.62	0.8	0.71	0.62	0.64
RMSE	0.79	1.02	0.91	0.81	0.84

Similar conclusions can be drawn when the HighFlow metric is used. Still, the use of these criteria could be of particular interest when the observed data is very scarce or uncertain, because the seasonal variation is still captured and their application is not dependent on a time step based comparison. Moreover, using time step based performance criteria could induce errors due to overfitting towards uncertain observations. However, for further analysis an improved general agreement is intended. So, the LowFlow and HighFlow metrics are left out and the combined criteria NSE-FDClow and NSE-FDChigh are used to evaluate the model results.

The performance metrics obtained from calibrations based on NSE only or its combined use (NSE-FDClow and NSE-FDChigh) are very similar. By imposing different weights to both parts of the combined criteria, more differentiation among the results can be obtained. Nevertheless, when comparing the resulting optimal parameter sets (Table 7.3), it can be seen that the similar performance stems from quite different parameter combinations. This indicates the problem of ill-identified parameters and confirms the identifiability issue, stating that different

parameter combinations can result in very similar model performance. At the same time, it indicates that optimizing towards these criteria alone is not sufficient to differentiate and identify the suitable parameter sets. A combined objective method leads to a more balanced model with respect to the calibration method used.

To study the model performance in more detail, Figure 7.2 shows the hydrological responses of the models with respect to the observed discharges and the results from the manual calibration performed by Vansteenkiste et al. (2011) for respectively three winter periods (winter 2003-2004, winter 2004-2005 and winter 2006-2007) and three summer periods (summer 2003, summer 2004 and summer 2006). The results of the NSE-FDChigh metric is shown for the summer period and the NSE-FDClow for the winter months, since emphasis lies respectively on high flows in summer and low flows in winter.

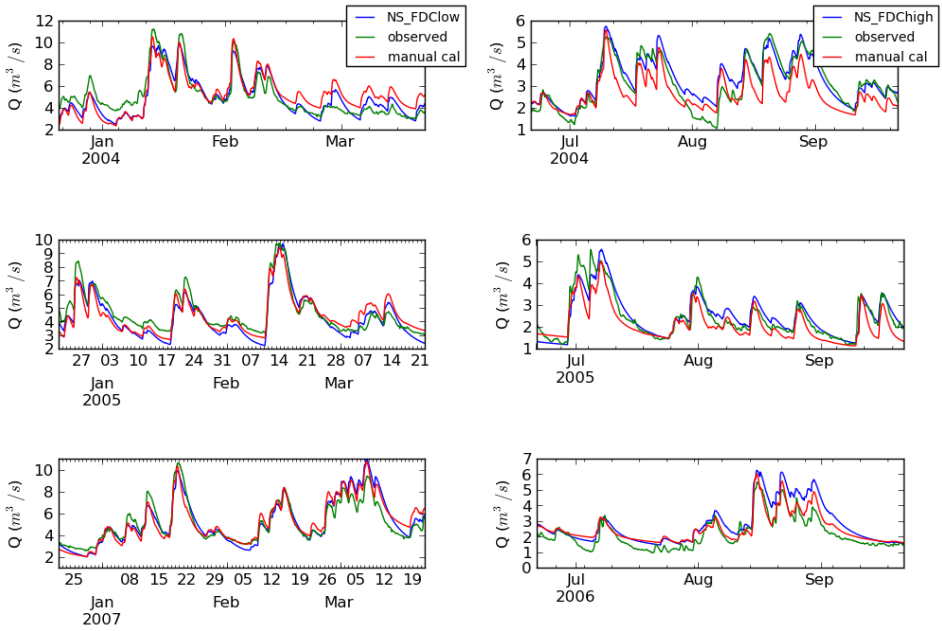


Figure 7.2: Observed, simulated by automated optimization and simulated by manual calibration of runoff discharges for winter events (left) and summer events (right) downstream the catchment based on the combined performance criteria. Selected winter and summer periods of the top two rows are within the calibration period (2003-2005) and of the last row within the validation period (2006-2008).

Table 7.3: Comparison of VHM calibrated parameters using the performance metrics NSE, NSE-FDClow and NSE-FDChigh. Notice that only a subset of the parameters presented in Table 6.1 is presented here, since it concerns only one structural option.

Parameters	NSE	NSE - FDClow	NSE - FDChigh	Manual cal
$u_{\max}$	200	346	418	220
u_{evap}	123	140	148	90
s_1	2.1	2.19	1.93	1.97
s_2	0.56	0.84	1.2	0.99
s_3	0.65	0.60	1.2	1.7
o_1	-3.97	-4.25	-3.84	-4.2
o_2	1.82	3.25	3.62	2.5
i_1	-3.23	-3.17	-3.49	-4.1
i_2	2.03	3.34	4.58	2.8
K_{g1}	2487	2500	2500	2100
K_{i}	150	150	150	120
K_{o1}	17	22	10	17
K_{o2}	41	30	78	17

As can be seen from Figure 7.2, the general evolution of the observed winter hydrographs is in good agreement with the simulated ones and in general very comparable with the results from the manual calibration. The recession limb of the simulated hydrographs matches the recession limb of the observed hydrographs, with more accuracy at the end of the winter by the automated calibration. Peak discharges seem to be well simulated by all models during the winter events. For the winters of 2003-2004 and 2004-2005 underestimations of the base flow are observed at the beginning of the winter periods, both by the automated and the manual calibrated models. Also during the validation period, as shown for the winter 2007-2008, the results are comparable.

For the summer discharges, more differences can be observed between the results coming from the automated and the manual calibration. The results from the automated calibration capture the recession limbs relatively better in the summers of 2004 and 2005, but overestimate the flows in 2006 (validation).

From the detailed graphical inspection of the discharges during some summers and winters it can be stated that the hydrographs simulated based on both calibration strategies show a satisfactory agreement with observed ones for winter as well as for summer periods. Both reproduce the small summer events as well

as complex-shape and long-lasting outflow regimes during and after winter events with acceptable matching capabilities.

It can be concluded that the chosen performance metric has a direct effect on the resulting optimal model realisation. The optimal realisation can be derived by optimizing a predefined metric or based on expert knowledge with a manual calibration. The results of the automated calibration procedure demonstrate the potential of applying automated calibration strategies as complementary procedure besides manual calibration. The validation results in Table 7.2 show that automatically derived parameter sets are also capable to reproduce the hydrograph outside the calibrated domain and that they are comparable to the validation results shown in Vansteenkiste et al. (2011). Accounting for the fact that these different regions in the parameter space are capable of reproducing the observed hydrographs, indicates the needed awareness for identifiability problems caused by overparameterization.

The time required to perform a manual calibration together with the inherent subjectivity hinder the reproducibility of a manual execution. Automatic calibration can be automated, reproduced and improved based on objective statements, making it a more robust scientific approach. Still, when using an automatic calibration, the choice of the parameter space should be well considered. Ranges should represent interpretable boundaries as much as possible. To combine the information from more performance metrics in the optimization of the model, the usage of a multi-objective calibration strategy could be chosen, searching for the optimal Pareto front instead of looking for one optimal combination (cfr. section 3.4.4).

7.3 Ensemble model calibration

In order to assess the relationship between the different model structures and different performance criteria, an optimization of the 24 rival model structures is performed to find the best performing parameter sets for each. Since manual calibration would be too time consuming and considering the subjectivity, the automated calibration procedure of the previous section was applied for each of the 24 rival models constructed in section 6.3.3.

Each model is calibrated for the NSE, NSE-FDClow and the NSE-FDChigh metrics separately over the calibration period with an initialisation period of 7 months to ensure that the results are independent of the initial conditions of the different reservoirs used. The model results of the warming-up period are ignored in the

computation of the different performance metrics. A maximum of 15000 model evaluations for each optimization was selected to limit the computational time, but convergence of the parameters was verified graphically when this limit was exceeded. The iteration towards convergence and the final converged values of the parameter sets are different between model structures, due to the different parameter interactions.

The aim of the comparison between model structures is to investigate how distinctive these performance criteria are towards the differentiation of the rival model structures and to evaluate the differences in parameterization obtained for the different model structures. To fully evaluate this effect, ranges of the parameter bounds were taken the same for all optimizations and similar for all model structures, provided in Table 6.1. Furthermore, the flow routing parameters are also calibrated, although it would be possible to derive these from the subflow filtering. By doing so, the discharge observation itself is the only used source of data.

Figure 7.3 shows the optimal values for the common parameters (i.e. parameters present in all model structures) for two different performance criteria for all 24 model structures. The parameter values are normalised based on the possible initial range given to the different parameters. As can be observed, most parameters vary in the entire range depending both on the performance criterion used as well as on the model structure.

It is noteworthy that all common parameters are included in all of the 24 model structures and representing similar 'physical' catchment properties in the different structures, except for the K_o value, which physical interpretation depends on the presence of an inter flow component in the model. Parameters coming from the linear and non-linear soil moisture storage component are expected to be different because of the different equations they are in, but based on the common conceptualization in the relation between water uptake and soil water content, both are considered in the analysis. As the task (conceptual representation) of these parameters is the same in the different model structures of the ensemble when optimized to a given performance metric, a similar value would be expected. However, the variation of the parameters amongst the individual model structures is striking. The different model structure combinations end up with alternative parameter combinations in order to optimize the performance criterion, driven by the optimisation algorithm.

The dependency on the performance metric has been reported earlier by Gupta et al. (1998) and Boyle et al. (2000), observing a similar effect when optimizing a single model structure on different time periods of the hydrograph.

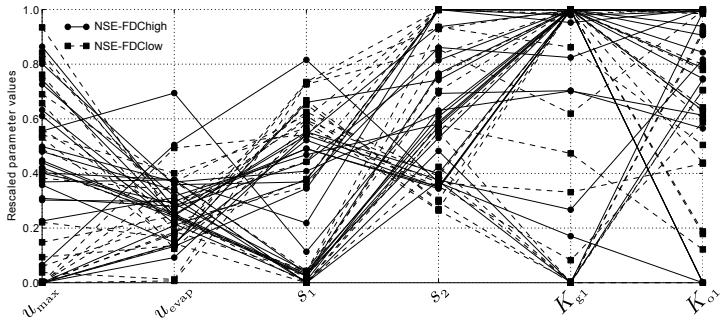
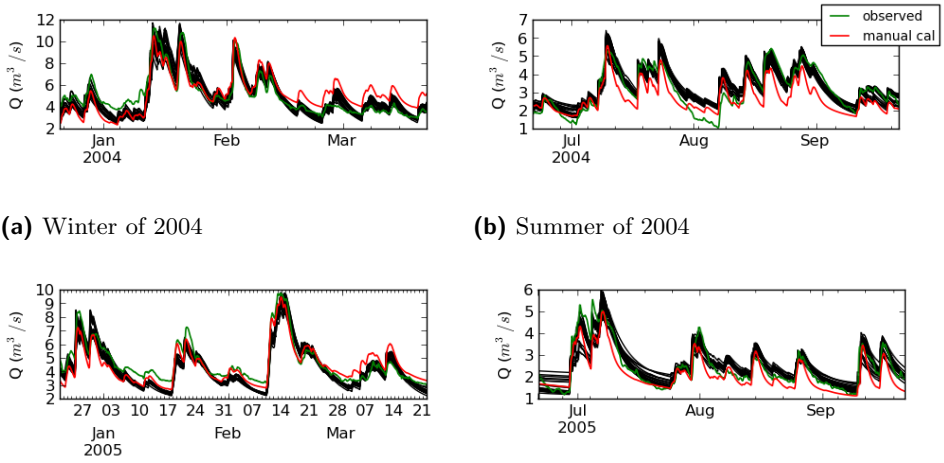


Figure 7.3: Overview of the derived optimal parameter values for the common parameters. The parameter sets of the 24 structures after optimizing towards the NSE-FDChigh(full line) and NSE-FDClow (dashed line) criterion are normalized towards their initially selected boundaries.

Differences in the resulting parameter values in a set of model structures when optimized to a common metric has less attention in the literature, partly due to the lack of flexible modelling environments available to do so (cfr. section 2.4.4). Similar studies focusing on model structures with common components and parameters such as Bai et al. (2009), Lee et al. (2005) and Clark et al. (2008) restrict their evaluation on the calculated performance metrics. Vaché and McDonnell (2006) compared the optimal parameter values of 4 model structures with common components. The resulting optimal parameter values were comparable for the most simple model structures with 3 or 4 parameters. However, similar to the results here, more complex model structures (in terms of number of parameters) resulted in less identifiable parameter values and more differences in the resulting optimal values.

Looking to specific events in Figure 7.4, it can be concluded that most of the time the ensemble is capturing the variations, but the models in this ensemble have a very similar representation. Hence, all of them are overestimating or underestimating the flow dynamics simultaneously and the effect of the individual model structure decisions is not resulting in distinct behaviour after the optimization towards a common performance metric. In other words, the structures are too interdependent to really represent completely different situations.

Confronting the similarity of the model output with Figure 7.3 leads to the conclusion that a high variety of parameter combinations realize a very similar behaviour amongst the individual members of the ensemble. Moreover, the manual calibration of structure *exp i a no*, added to Figure 7.4, is more distinct then the members of the model structures. Hence, the degrees of freedom (parameters) of the model



(a) Winter of 2004

(b) Summer of 2004

(c) Winter of 2005

(d) Summer of 2005

Figure 7.4: Resulting predicted flow downstream the catchment of the optimal model realisations for the winter (left) and summer (right) of two consecutive years of the calibration period. The results of the ensemble of 24 model structures after optimization are added as black lines. The optimal realisation of the manual calibration of structure *exp i a no* (red) line is more distinct as the realisations by the individual members of the ensemble.

structures are more dominant to change the model behaviour as the model structure variations. This indicates a lack of identifiability of the model structure itself, which hampers the identification of better model structural decisions. Hence, taking into account more differentiation in the model structural hypotheses and the identifiability of the individual model structures is crucial to be able to identify the suitable model structural options based on optimization.

Since, based on the analysis, no specific model structure outperforms the other structures it could be questioned how to use these results. One possibility is to combine the output of the different structures. The most straightforward approach is to take the mean of the individual structures assuming they are equally reliable, but more elaborated methods for multi model ensemble analysis exist and could be applied, e.g. Bayesian Model Averaging (BMA) (Vrugt and Robinson, 2007) or probabilistic analysis of the individual structures towards specific signatures (Georgakakos et al., 2004). These methods are out of scope for this dissertation, but they provide a working solution for ensembles of models.

Taking into account the mean of the ensemble, as said a rather conservative working solution, the performance can be evaluated using the known performance met-

rics. Looking at Table 7.4, the performance metrics of the ensemble average is comparable to the measures obtained by the manual and automated calibration.

Table 7.4: Performance values for the calibration period (2003-2005) when using the mean of the ensemble of 24 structures

	Mean of ensemble performance on calibration period
NSE	0.86
R^2	0.93
MAE	0.56
RMSE	0.77

7.4 Conclusions

In this chapter, the evaluation of the model structure decisions defined in chapter 6 is attempted by optimizing the ensemble of model structures. The aim is to identify which model decision can outperform the others.

In the first part of this chapter the relationship between the used performance metric and the calibrated parameter values for one specific model structure is tested. Based on a set of performance metrics with a specific focus on low and high flow, optimal parameters were derived for each and the outcome of the automatic calibration was found comparable to the manual calibration (Vansteenkiste et al., 2011), in terms of performance as well as for modelling specific summer and winter events. However, different parameterizations are obtained when optimizing towards different performance criteria (NSE and separate criteria focusing on high and low flows), indicating a lack of identifiability of the model structure. These results are in line with the work presented by other authors when evaluating a single model structure (Gupta et al., 1998; Beven, 2008b).

This confirms that using this type of (non identifiable) lumped hydrological model structures, which is common practice in both operational and scientific application, the decision of the performance metric should be in direct correspondence with the model purpose. The direct relation with the performance metric means that the model is only valid for the specific purpose inherited in the constructed performance metric. This limits the predictive applicability of the model and this limitation should be clearly communicated.

In the second part, the optimization is extended to the ensemble of model structures defined in chapter 6 and based on 4 model decisions. In contrast to the evaluation of a single model structure, an evaluation of the model parameters that are shared by an ensemble of model structures is rather exceptional in literature.

For the model ensemble and the optimization applied in this study, it could be concluded that the performance of the individual model structures in the ensemble is comparable. For the defined VHM alternatives, no model structure outperforms the other model structures and the representation is highly similar. The conceptual differences provided by the alternative model decisions could not be distinguished in the optimal realisations of model structures.

Furthermore, the contributing parameter values have a striking variation amongst the model structures, nevertheless their common function in the model structure. The conceptual function of the common model parameters within the ensemble is expected to be the same when optimized to the same performance metric. Apparently, the degrees of freedom (parameters) of the individual structures are more decisive than the structural differences in order to differentiate them from each other. This is probably due to parameter interactions leading to multiple combinations that are able to provide a similar performance, i.e. a lack of parameter identifiability.

In summary, the results of this chapter indicate the lack of identifiability (each individual) and a lack of differentiation (comparing them) amongst the different structures of the used ensemble. These results are based for the defined set of model alternatives of VHM and further studies are needed in order to generalize these statements.

Further analysis would also benefit from more distinctive model structural hypotheses. At the same time, when aiming for a process-based model structural comparison, where systematically single components are interchanged and compared, a limited ability to distinguish model structures is expected. Hence, the impossibility to distinguish will probably hamper a optimization based strategy. The latter is the starting point for the next chapter, where a new model structural comparison technique is developed focusing on the comparison of model structures with a major number of corresponding components but without being dependent on the model performance itself.

CHAPTER 8

A qualitative model structure sensitivity analysis method to support model selection

Redrafted from

Van Hoey, S., Seuntjens, P., van der Kwast, J., and Nopens, I. (2014b). A qualitative model structure sensitivity analysis method to support model selection. *Journal of Hydrology*, 519:3426–3435

8.1 Introduction

A flexible approach of model construction, as it was implemented for the VHM model in chapter 6, enables an increased ability to compare and test different model structures, each representing a different set of assumptions. In the previous chapter 7, each of the model structures was calibrated for both low and high flow performance metrics. The result confirms the lack of identifiability in the parameter sets as well as the dependency of the retrieved optimal parameter values on the used performance metric (Gupta et al., 1998). However, the identifiability problem is not tackled by the optimization itself. Moreover, due to the high similarity of the different model structures, optimization towards the used model structures is not sufficient to distinguish them. The latter makes it insufficient to guide us in the model identification process.

In this chapter, instead of comparing the different model structures with respect to their performance, sensitivity analysis is used to guide the model selection

within the set of model structures. Assessing parameter sensitivity is used regularly to identify non-influential model parameters towards a chosen (aggregated) model output. We explore the applicability of using a sensitivity analysis on model structures rather than model parameters. In chapter 3 different methods for global sensitivity analysis were presented and implemented. In analogy with parameter sensitivity analysis, evaluating the effect of certain model structure components could reveal the added value of the component towards specific performance metrics and as such, assist in model selection.

To do so, we have to assume that the effect of a model component can be evaluated based on the change in parameter influences. In short, a change in a specific component results in changing parameter influences towards the performance metric, the adaptation leading to this change in sensitivity is considered to give the model configuration added value (i.e. predictive performance). This chapter introduces a component-sensitivity concept in a qualitative (graph-based) manner.

The component-based sensitivity analysis is first introduced and the results are discussed in this chapter. The methodology is presented for the set of model variations of the VHM structures implemented in chapter 6. By comparing the effect of changes in model structure for different model objectives, model selection can be better evaluated.

8.2 Extending parameter sensitivity towards model component based sensitivity analysis

The presented methodology is a direct extension of the Morris screening approach from chapter 3, applied on multiple model structures with partly similar components. The following steps need to be taken:

1. *Decide about model parameter distribution for all parameters of all model components taken into account*:
   As in all global sensitivity analysis methods, the distribution of all the parameter values needs to be chosen in order to evaluate the effect of the parameters for different parameter sets. As clarified in section 3.5, it basically comes down to sampling uniform in the $[0-1]$ range and using a proper inverse CDF fucntion. Similar to other sensitivity methodologies, the decided parameter ranges will influence the results and must be chosen carefully.

2. *Perform Morris screening for each model structure*:
    - Derive the optimal parameter trajectories, according to Campolongo et al. (2007)
    - Run the model $r(k+1)$ times for each parameter set of the different trajectories, with r the number of optimal trajectories and k the number of parameters
    - Calculate the μ^* value for each objective function or output variable considered
3. *Visualize the change of parameter influence for each model decision*:
    - Split the set of μ^* values in 2 groups, according to the two different model structure variations
    - Make a scatter plot of the μ^* values and add the 1:1 line (bisector) to create the evaluation chart (Figure 8.1)

As such, the structural alternative of the OAT method is derived. Every plot compares the outputs of the variation in one specific model component, while other components are in both cases equal. In other words, the deviation from the 1:1 line is due to the change of that component. However, different deviations (i.e. changing the same component starting from different model setups) are plotted together to check for recurrent sensitivity effects caused by the difference in the specific model component. This is similar to an OAT approach, where the effect of a parameter is evaluated by combining the output from different runs, each with a different initial starting point in the parameter space.

The obtained evaluation chart, as shown in Figure 8.1, can be interpreted based on two criteria: (d_1) The distance from the origin relates to the relative importance of the parameter and (d_2) the perpendicular distance from the bisector indicates the parameter influence deviation introduced by the model adaptation. To assist in the interpretation of these type of figures, 4 different zones are indicated, termed X_1 to X_4:

- type X_1: Parameters used in one model option and not in the other appear with their influence on the x- or y-axis. The distance to the origin is related to the influence of the parameters. High values mean that these parameters have a major influence on the output variable and as such, influence the output variable considerably.
- type X_2: Parameters present in both model options, but with no major change in parameter influence. This means the change in model compo-

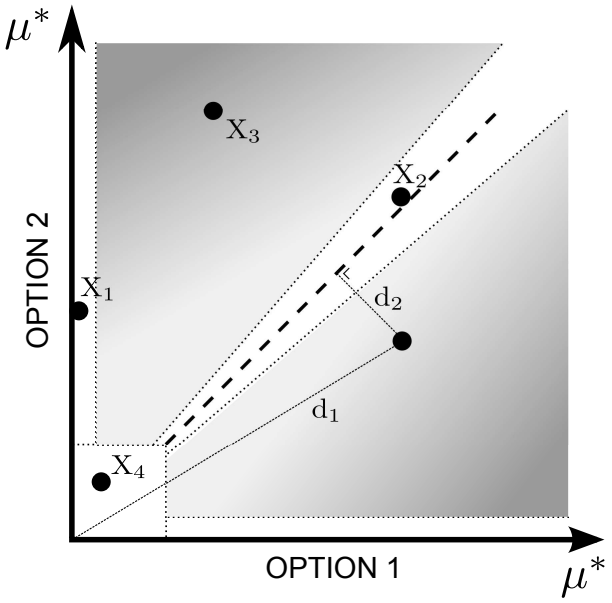


Figure 8.1: Illustration of the model structure sensitivity evaluation chart, defining the four different zones that characterise the parameter influence X_1 to X_4 and the two major criteria d_1 and d_2 are defined. Mainly parameters corresponding to zone X_3 do indicate impact on the structural sensitivity.

nent has no influence on the way the parameter influences the output variable (probably coming from another component). Mainly for larger μ^* values, these parameters indicate large influence and no interaction with either model options (conditions for identifiable parameters).

- type X_3: The combination of a large bisector deviation and a large μ^* value (dark gray) is typical for parameters mainly influenced by the model option. If the parameter belongs to both model components, the option with the highest μ^* leads to increased influences towards the output. For parameters not belonging to the model option components, the degree of parameter interaction is related to a shift in the influence of that parameter.

- type X_4: Low μ^* values in both model options are related to non-influencing parameters towards the (performance) metric considered, suggesting potential overparameterization and room for model reduction.

The method assumes that the shift in parameter influence indicates the effect of the model component it represents. However, parameter interactions are present in most models and parameters are affecting other components behaviour. In the evaluation chart, this effect is visualised by parameter shifts along the x or y axis of parameters not included in one of the two model options. As such, the methodology uses an implicit way of evaluating parameter interactions for those parameters.

The assumption that the parameters are representative for the behaviour of the model component they are part of, is used for their representation in the evaluation chart. All parameters are given a component identifier, based on the equation that the parameter is part and the model structure component that the equation is part of. Five model components are identified for the model structures of the VHM model: (1) the storage component, S, which refers to the linear or non-linear storage component; (2) evapotranspiration, E, which is the same for all model structures in the ensemble, i.e. linear relationship with the model storage; (3) overland flow component, O; (4) interflow component, I and (5) the baseflow component, B. These identifiers are also added as a column in Table 6.1 and plotted as such in the evaluation graph (see further).

8.3 Results

8.3.1 Parametric sensitivity analysis

The proposed methodology includes the execution of a Morris screening for each model structure. For each of the applications, a subset of 20 trajectories was selected out of an initial sample of 500 trajectories, maximizing the distance between the pair of trajectories (Campolongo et al., 2007; Saltelli et al., 2008). Furthermore, a visual control of the histograms was done to ensure the frequency of the different levels was comparable. A uniform distribution for all parameters was assumed and discretized in $p = 4$ levels as suggested by Morris (1991), the ranges of sampling are given in Table 6.1.

The sensitivities based on the Morris screening are shown in Figure 8.2 for one specific model structure, i.e. the *exp ia no* (see Figure 6.11 for the component option labels), estimated for the performance metrics NSE, NSE-FDClow and NSE-FDChigh. The structure has a non-linear storage, both interflow and antecedent rain included, but no extra routing complexity. For every parameter,

sensitivity indicators μ^*, μ and σ are computed and the combination yields information on the relative importance of the parameters and the amount of interaction between different parameters. Background on how these sensitivity indices can be interpreted is given in section 5.3.

Figure 8.2 indicates a different parameter sensitivity behaviour when using different metrics. Similar absolute values for both μ^* (mean of the absolute values of the EEs) and μ (mean of the EEs) in combination with an opposite sign as in Figure 8.2b and Figure 8.2c for parameter K_{o1} means that the parameter inversely influences the chosen metric. In other words, smaller values for parameter K_{o1} will increase the performance metric (improved performance). This corresponds to the result of the optimized parameter set in Table 7.3, where K_{o1} is 22 and 10 when using respectively the NSE-FDClow and NSE-FDChigh metric. The defined range for K_{o1} is 10 till 120 (Table 6.1). When μ is low and μ^* high, the parameter has a large effect on the chosen metric. However, the high σ values are a clear indication of the interdependence between the parameters, making it difficult to directly link parameter influence on the chosen metric. The main reason for the interaction is the translation of the conceptual model (all flows are fractions of the incoming rainfall) to the mathematical model. When fractions are calculated for a single time step, the sum of individually calculated fractions can be larger than 1. Hence, the fractions need to be rescaled before the corresponding flux is calculated in order to keep conservation of mass.

The relative differences in the parameter influences towards varying performance metrics are caused by either their influence towards the different aspects of the hydrograph or a change in interactions between the parameters towards these performance metrics. For all three criteria tested, the soil storage parameter s_1 is of major importance. When focusing on NSE or specifically on low flow, the inter flow parameters i_1, i_2 and K_i have increased influence indicating the importance of the inter flow compartment to allow describing the low flow variability. When focusing only on high flows, the variability in the peak flows is mainly explained by the storage parameter s_1 and the routing of the overland compartment K_{o1}.

8.3.2 Component sensitivity analysis

The evaluation graphs of all model combinations are shown in Figures 8.3, 8.4 and 8.5 for the NSE, NSE-FDClow and NSE-FDChigh, respectively as introduced in section 6.4.

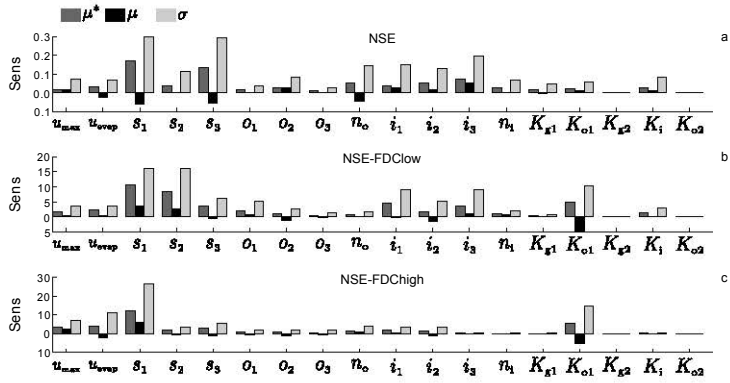


Figure 8.2: Morris sensitivity screening method evaluation of the structure with non-linear storage component, interflow and excess infiltration (*exp ia no*) for the three different objectives: NSE (a), NSE-FDClow (b) and NSE-FDChigh (c). Three indices are plotted for each parameter: μ^* in dark grey, μ in black and σ in light grey.

The values are the μ^* values, giving information about the relative importance of the parameters as it represents a good proxy for the total variance. The labels of the parameters are given the first character of the component they belong to, as shown in Table 6.1. This enables to quickly see which components are contributing the most towards the variation in the output and how these sensitivities are changing when adapting the model structure.

Each graph consists of a number of points equal to the number of parameters k multiplied with the number of compared model structures. The latter is equal to 12 in the case of a the linear storage, inter flow and infiltration excess structural decision (1 on 1). In the case of the routing decision, the added complexity is combined in a single plot, resulting in twice the comparisons between 8 structures.

Since the Elementary Effects provide qualitative information, only the relative values of the μ^* are interpretable and the differences in absolute values among the figures are irrelevant to compare. The adaptation of the font-size is mainly to improve readability of the component labels.

When using the NSE, shown in Figure 8.3, the parameters of the storage component (S) present in the non-linear storage component are most influential to the NSE performance metric. When changing the storage component from a non-linear to a linear component, these highly influential parameters are no longer included in the model (type X_1). However, this model structure change gives rise to an increased influence of mainly the overland flow parameters (type X_3). In other

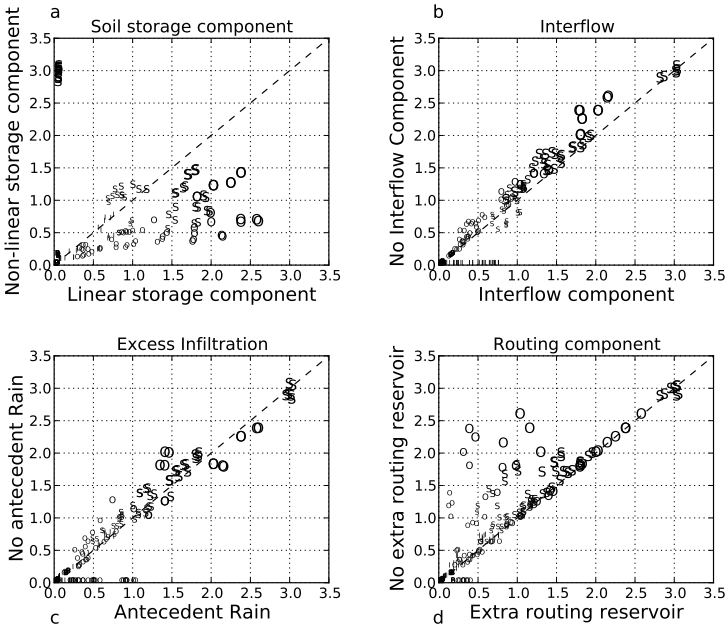


Figure 8.3: Effect of the selected model structure on the sensitivity of the NSE efficiency. Each subplot compares the variation of one structural component while keeping the other components fixed. The characters are the μ^* value, representing the model component the parameter belongs to (O = Overland flow, S=Soil storage, I=Interflow, B=Baseflow and E= Evaporation parameters).

words, the hydrograph is mainly fitted by the overland parameters in the linear case, whereas in the non-linear case they are fitted by these non-linear storage parameters. Considering the fact that the linear model has less degrees of freedom (parameters) and an increased sensitivity, the non-linear parameterization potentially leads to overfitting.

In general, inter flow parameters are not very influential to the NSE metric, as evidenced by the low μ^* values. Excluding the inter flow component results in a slightly higher influence of the overland and storage parameters, but the effect is less pronounced compared to the case including the storage component (type X_2). The inter flow component could potentially be excluded as a model reduction step. Adding complexity with an excess infiltration component gives comparable results and excluding the extra routing reservoirs results in a major increase of the influence of the included model parameters (type X_3), suggesting the model

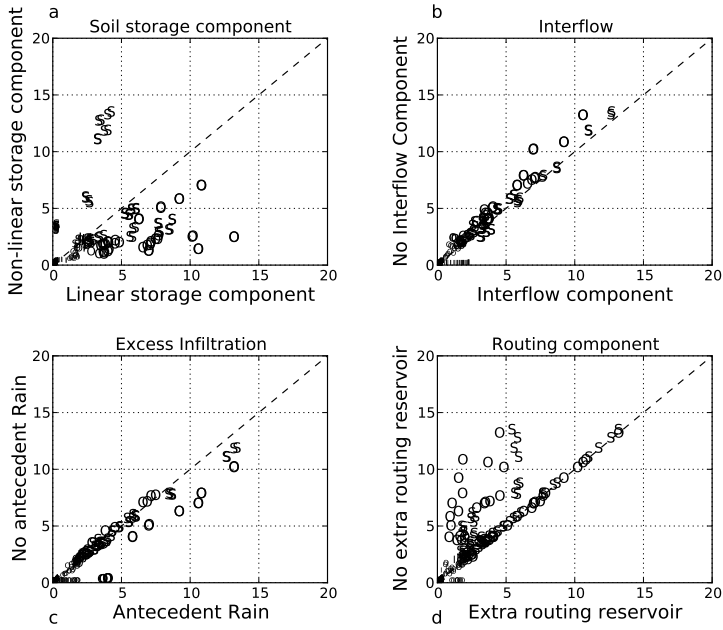


Figure 8.4: Effect of the selected model structure on the sensitivity of the high flow criterion. Each subplot compares the variation of one structural component while keeping the other components fixed. The letters are the μ^* value, representing the model component the parameter belongs to (O = Overland flow, S=Soil storage, I=Interflow, B=Baseflow and E= Evaporation parameters).

can be simplified to reproduce the hydrograph based on the NSE performance metric.

Adding a routing component affects also the influence of non-routing parameters, visualising the effect of parameter interactions in the model. Furthermore, base flow and evapotranspiration parameters are in general not very influential towards the NSE performance metric.

Focusing only on high flow, as in Figure 8.4, gives very similar effects, but the dominant effect of the storage parameters is less apparent compared to the NSE case. Again, a shift from influential overland flow parameters towards storage parameters when going from a linear to a non-linear component is visible. Interestingly, the non-linear storage parameters of Figure 8.3 are yet not very influential (low values of type X_1). Instead, the shift occurs in the common storage parameters. As such, the selection of the linear or non-linear component is a decision between

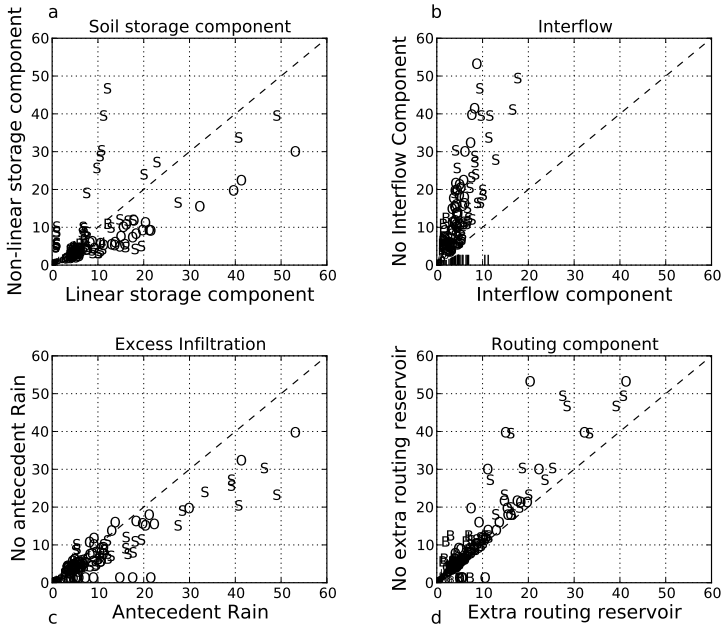


Figure 8.5: Effect of the selected model structure on the sensitivity of the low flow criterion. Each subplot compares the variation of one structural component while keeping the other components fixed. The letters are the μ^* value, representing the model component the parameter belongs to (O = Overland flow, S=Soil storage, I=Interflow, B=Baseflow and E= Evaporation parameters).

giving the influence to the storage parameters (changing internal catchment state) or the overland parameters (changing lag times). Since in general, the aim is to find parsimonious model structures with identifiable parameters, an increased influence of the model component itself is preferred. As such, the non-linear case is recommended in this case. Again, the addition of extra routing reservoirs is only decreasing the influence of the parameters of both overland and storage.

The similar observed effects of the sensitivity shifts between Figures 8.3 and 8.4 can be explained by the focus both metrics are giving to high flow periods. This proves at the same time the stability of the sensitivity measures, indicating that sufficient sample trajectories have been used.

The influence of the parameters towards the low flow criterion is given in Figure 8.5. The storage and overland flow parameters are dominating the variation in the

output. Concerning the soil storage component, a similar behaviour as the high flow objective function is observed, with parameters of type X_3.

The major difference with the previous objective functions is the shifts for both inter flow and excess infiltration. The addition or removal of the inter flow component has a much larger effect on the influence of the different parameters for the low flow oriented metric compared to the other metrics. This confirms the relation between model structure properties and performance criteria used. Although one could expect the incorporation of an inter flow component would result in high sensitivities for this low flow criterion, it is not the case. The large shift in sensitivity between both options (type X_3) does confirm the relative importance of the model component, the more simplified model (no inter flow component) results in more influential parameters for overland and storage parameters probably due to parameter interactions.

For excess infiltration, adding extra complexity does result in higher influence for the parameters. This could potentially be linked with the effect of a sudden rain event during a dry period. The excess infiltration adds the possibility of generating overland flow, giving the model a quick response time. Without this component, the soil storage needs to be filled before runoff takes place. Further performance checks during different phases of the hydrograph could confirm or reject this hypothesis. Evaluating the model performance during selected storm events characterised by intense rainfall intensities after a dry period would be a good starting point to do so.

8.4 Discussion

Based on the presented application of the component sensitivity analysis methodology on the used flexible model structure under study, several suggestions can be made with regard to model selection: (1) A non-linear storage component is recommended, since it ensures more influential (identifiable) parameters for this component and less parameter interaction; (2) Interflow is mainly important for the low flow criteria; (3) Excess infiltration process is most influencing when focussing on the lower flows; (4) A more simple routing component is advisable; (5) Baseflow parameters have in general low influence, except for the low flow criteria. Furthermore, based on the comparison of the used objective functions, it can be stated that a more simple model is able to reproduce the hydrograph when the focus is on high flows. When the goal is to take into account the low flows as well, a more elaborate model description is required.

These advices can be derived for the used case study without executing any model optimization algorithm, but merely by screening the parameter space with a global sensitivity analysis. The methodology depends on a set of Morris screening outputs. The results are brought together in an evaluation chart giving a qualitative assessment of the model structure decisions. The number of model runs depends on the selected set of trajectories (here 20 was used) and the number of parameters in each model structure (between 8 and 16), giving a total of 180 and 340 runs for each model.

However, some assumptions are taken when performing the analysis. It is assumed that the shift in parameter influence is an indicator for model component importance. This is however limited to the specific aggregated or performance metric used and cannot be extrapolated towards other metrics as the example analysis has shown. Parameter dependencies are treated implicitly, when parameter shifts (i.e. change in sensitivity of the parameter when switching the model option) are occurring for parameters of components that are not in either of the two model options. The sensitivity of these parameters (not part of any of the two model options to compare) changes due to the interaction effects they have with other parameters. That is the reason why the output of the σ values are not included in the evaluation graphs.

The evaluation of the shift in sensitivity when changing model components gives added value to a traditional sensitivity analysis of a single model structure, mainly in comparing multiple model structural options in a flexible environment.

By comparing the differences in sensitivities between structures, those components with the most potential for improvement can be identified. Hence, these model components are characterized by parameters with low influence, meaning that alternative configurations for these components may give rise to the largest differences.

The dependence of the outcome of the method on the performance metric used, confirms again the importance of selecting the appropriate metric in order to converge to a suitable model structure. With the current availability of frameworks for flexible model development, both in hydrology (Clark et al., 2008; Fenicia et al., 2011) and other fields (Wesseling et al., 1996), methods for model evaluation and comparison as the one presented in this chapter are essential to really benefit from the flexibility in model building.

8.5 Conclusions

The use of a flexible model structure provides a useful way of testing and comparing different model structure hypotheses but requires dedicated methods for model selection and comparison. A straightforward method is presented to support this process of model evaluation when different rival models with overlapping components are available.

The method directly builds on an existing global sensitivity analysis technique (i.e. Morris screening). Applying it to multiple models simultaneously and bringing together the information in a single evaluation chart, allows performing a qualitative evaluation of the different model options tested based on the shifts in sensitivity.

The used performance metric can be selected in function of the specific application and is not limited to the ones presented in this chapter, making it fit with the metric-oriented approach described earlier. As illustrated by the application on a set of 24 model structures, the information extracted is useful for model selection in relation to the used performance metric. The proposed evaluation method is generic and can also be applied to models in other scientific fields than hydrological modelling.

PART IV

DIAGNOSING STRUCTURAL ERRORS IN LUMPED HYDROLOGICAL MODELS

CHAPTER 9

Communicating lumped hydrological model structures: a Gujer matrix analog

9.1 Introduction

The focus of lumped hydrological models is to represent the dominant processes of water within a catchment by representing the catchment as a set of connected reservoirs, excluding spatial heterogeneity. These models are both used in operational settings (forecasting, integrated modelling) as well as a research tool to understand and get more insight in the system functioning, since they provide descriptions at a low computational cost (Clark and Kavetski, 2010; Wagener et al., 2001b).

A wide range of model codes and alternative implementations are available to develop lumped hydrological models, developed during several decades (for a historical background of these reservoir type hydrological model structures, the reader is referred to Beven (2012), p36). Some of them are highly popular and applied frequently in literature, whereas the reasoning to select a specific model structure is not always clear (apart from its availability reason) and the implementation not always available. The abbreviations of some well-known models are commonly used terminology within the hydrological modelling community (BHV , SAC-SMA, PDM, NAM, HYMOD, GR4J, TOPMODEL, VIC...), but some limitations can be identified which are not in line with the requirements as defined in section 2.5.2:

- **Flexibility is limited** for these model structure definitions. The user mostly has to choose the model as such or choose another model. However, the uniqueness of each model study could require a mixture of existing model concepts. The model structure sensitivity analysis approach explained in chapter 8 would not be feasible on a wide range of model structures, since the possible adaptations of individual model process components are restricted. To be able to test different process descriptions and change processes one at a time, the model architecture needs to be flexible.
- The **mathematical and computational model** of these structures are regularly **not separable**. The ability to define the mathematical and computational model independently is defined as a central requirement in section 2.5.2 to support model evaluation. Although the numerical solution is recognized as a critical step in the model building process (Beven, 2001), numerical time stepping schemes for hydrological models have received surprisingly little attention in the hydrological modelling literature.
- The specified acronym cannot always be linked to one specific implementation of the model due to a **lack of transparency of the source code**. This hampers the comparison of different model outputs as these differences can originate from the model structure conceptualisation, but as well from the implementation itself.

Recent research provides a more general description of lumped hydrological models, enabling the definition of different model conceptualizations within a generic flexible framework (Clark et al., 2008; Fenicia et al., 2011; Kavetski and Fenicia, 2011). In practice, these flexible modelling environments actually boil down to the definition of these models as a set of ODEs. Indeed, lumped hydrological reservoir models do not differ from the general mathematical formulation of Equation 2.1 (section 2.2) and can be formulated as a set of ODEs by defining the appropriate mass balances.

This chapter starts from the interpretation of lumped hydrological models as a set of ODEs to overcome the limitations enlisted above. The aim is to propose a way to easily communicate about lumped hydrological models independently from the implementation (source code) itself while supporting maximal flexibility in the chosen model structure configuration.

To accomplish this, a method to communicate about these lumped hydrological model structures in a standardised way is proposed, i.e. by summarizing the model structure in a single matrix representation. It is inspired by similar representations used in (bio)chemical research, commonly seen in a wastewater treatment mod-

elling context (Gujer and Larsen, 1995) and adapted for pharmaceutical processes as well (Sin et al., 2008). This representation enables to communicate about model structures in a standardised and transparent way, supporting more transparent and reproducible scientific reporting.

The remainder of the chapter is structured as follows. First, a short introduction about existing flexible frameworks for lumped hydrological modelling is given, illustrating that maximal flexibility is provided by defining a set of ODEs. Next, the motivation to propose a standardised representation rather than yet another implementation is discussed. The original Gujer matrix representation is shortly introduced, after which the hydrological variant is explained in detail. In the last part, the representation is applied on a number of existing lumped hydrological models.

9.2 Flexibility of lumped hydrological model structures

Flexible environments do exist for hydrological modelling. Kralisch et al. (2005) illustrates how general purpose flexible model environments can be used to develop and apply hydrological models, which practically means that one has to implement new components in scripting language. The latter is similar to the usage of a domain specific programming language for catchment modelling as it has been developed for distributed modelling (Kraft et al., 2011; Kraft, 2012; Schmitz et al., 2013). The scripting based approach provides ultimate flexibility, but the model structures that can be implemented in flexible model environments like in Wagener et al. (2001a); Clark et al. (2008) and Fenicia et al. (2011) are focusing specifically on hydrology. They can be summarized by the combination of a soil moisture accounting module and a routing module, where different options can be selected for both parts. Similarly, the Hydromad package in R, developed by Andrews et al. (2011) and inspired by the PRMT package of Wagener et al. (2001a), also provides multiple options of existing models for both a soil moisture module and a routing module. Bai et al. (2009) uses a modular modelling structure of three modules: Soil moisture accounting, actual evapotranspiration and routing, with different options for the three components.

All of these model environments act as a container to interchange existing models and keeping the comparison on a rather coarse granularity for interchange (section 2.5.2). They do not provide direct interchanges on process level and lack a unified framework. In this respect, the flexible approach formulated by Fenicia

et al. (2011) and Kavetski and Fenicia (2011) break down the hydrological structures in reservoir, lag and junction elements that can be recombined to build new model structures and supporting flexibility on a fine granularity which enables the evaluation of individual model components.

Similarly, the concept of Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008) is of interest, since it combines individual modelling options from well-known hydrological models to construct new equally plausible model structures, where the model components can be evaluated separately. To accomplish this, the framework translated existing models as a set of ODEs. Indeed, despite the impression of large distinctions made by different naming and descriptions, most of these models share a similar underlying framework of connected reservoirs and are all based on ODEs, convertible to the general model layout given in Equation 2.1.

The work of Clark et al. (2008) and Fenicia et al. (2011) illustrates that existing lumped hydrological model structures can be expressed as a set of ODEs. Hence, when looking from a system dynamics approach, the required flexibility is achieved when direct insight is given into the equations itself. When doing so, individual components (equations, processes, fluxes...) can be adapted whilst keeping the other elements fixed to enable model comparison on a process level.

Moreover, the definition of a model structure by a set of ODEs enables a separate definition of the mathematical and computational model. Using continuous time for the model formulation and approximating it in discrete time to solve the model numerically provides the flexibility in changing the model step size and choose the most appropriate numerical solver (Clark and Kavetski, 2010; Kavetski and Clark, 2011).

9.3 Standardisation of model structures

Developments such as the FUSE environment (Clark et al., 2008) and SUPERFLEX (Fenicia et al., 2011; Kavetski and Fenicia, 2011) make a system dynamics approach of existing lumped hydrological modelling possible. Although the FUSE implementation compiles a great set of existing model structures, the possibilities are still rather limited from a hypothesis testing point of view, being limited to a two-layer configuration. The flexible approach proposed by Fenicia et al. (2011) and Kavetski and Fenicia (2011) enable a further generalisation in model structure construction by using reservoir elements, lag functions and junction elements as basic building blocks to represent different flow configurations.

Still, both FUSE and SUPERFLEX provide a direct construction of model structures as a set of (non-linear) ODEs. As stressed by Fenicia et al. (2011) themselves, the focus is not on a particular computer code or on software design aspects, but on the conceptual elements supporting controlled flexibility in hydrological models. FUSE provides a fast interface for hydrological modellers to create and test a variety of existing model structures and it illustrates as such the similarity in the mathematical foundation of most of these models (proof of concept). However, in essence it just adds a domain-level layer on top of general ODE implementations, as is done in other scientific fields like water quality modelling, ecological modelling or chemical engineering.

Existing lumped hydrological model structures such as PDM discuss alternative model structure options as well (Moore, 1985), providing them some level of flexibility (mostly just depending on the available implementation or software). However, in most cases, the authors just mention the PDM model acronym referring to the name but giving little insight in the specific options used, which can differ between implementations and between different research institutes.

The lack of transparency about model structure implementation is currently more of a problem than the availability of model software environments. Hence, easy and interpretable communication of the chosen model structure is essential to ensure that the implementation can be done in any environment or software most suitable for the user. This chosen model structure can be any of the legacy models, a configuration of the FUSE or SUPERFLEX environment or any newly defined model structure. By providing a way to communicate about the model structure in a generic way, the modeller has maximal flexibility in the (software) tool used. Hence, to improve the communication and reproducibility of scientific publications on this topic, focus should go to a standardised approach to communicate about model structure decisions.

The combination of the ODE representation in a matrix representation and the description of the used numerical solver (ideally, an open source implementation) provides the necessary information to communicate about any model structure configuration in a reproducible way, independent of a specific software environment.

9.4 The Gujer matrix representation

Standardisation of model structure definitions has been used in different disciplines, such as waste water treatment modelling (Gujer and Larsen, 1995) and

pharmaceutical modelling (Sin et al., 2008). The International Water Association (IWA) task group on *Mathematical modelling for design and operation of biological waste water treatment* introduced a model representation for biokinetic models such as the ASM family (Henze et al., 1983; Gujer and Larsen, 1995).

Table 9.1: Standard representation as a Gujer matrix of a process model consisting of state variables $S_1,\ldots,S_m$, processes $p_1,\ldots,p_n$, stoichiometric coefficients $\nu_{1,1},\ldots,\nu_{n,m}$ and kinetics $\rho_1,\ldots,\rho_n$

	→ continuity					
mass balance ↓	process p_i	components				reaction rate ρ_i
		S_1	S_2	…	S_m	
	p_1	$\nu_{1,1}$	$\nu_{1,2}$	$\cdots$	$\nu_{1,m}$	ρ_1
	p_2	$\nu_{2,1}$	$\nu_{2,2}$	$\cdots$	$\nu_{2,m}$	ρ_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	p_n	$\nu_{n,1}$	$\nu_{n,2}$	$\cdots$	$\nu_{n,m}$	ρ_n
	stochiometric parameters	S_1 full name	S_2 full name	…	S_m full name	kinetic parameters

When applied to (bio)chemical reactions, a process is described by a reaction rate and by the stoichiometric coefficients for all components involved in the process. The mass balances for all components are described by a set of ODEs, taking into account the stoichiometry, the reaction rate and the sign of the reaction (production versus consumption), all summarized in a matrix representation (Table 9.1). The matrix is composed of the following elements:

- the left column lists all n processes p_i accounted for in the model
- the top row lists all the m different components S_j taking part in the processes
- the right most column lists the reaction rates ρ_i for the respective processes in the left column
- the core part of the matrix represents the stoichiometric coefficients $\nu_{i,j}$
- the left bottom cell lists the stoichiometric (occurring in the matrix core cells) parameters, the right bottom cell lists the kinetic (occurring in the right column) and the center bottom cell the component full names

As such, the total transformation rate of a component S_j is given by

$$r_{\mathrm{j}} = \sum_{i=1}^{n} \nu_{\mathrm{i,j}} \rho_{\mathrm{i}} \quad j = 1, \ldots, m \tag{9.1}$$

where r_{j} is the total transformation rate of the component $\mathrm{S_j}$, $\nu_{\mathrm{i,j}}$ is the stoichiometric coefficient of the substance $\mathrm{S_j}$ for the process p_{i} and ρ_{i} is the reaction rate of the process p_{i}. Non-zero elements of a row of the matrix represent which components are affected by a given process. In other words, non-zero elements of a column indicate which processes have an influence on a given component or in which processes the component takes part. The signs of the stoichiometric coefficients indicates consumption (-) or production (+) of the corresponding component.

An example of the matrix representation was provided in section 4.2.2 to clarify the respirometry model used (Table 9.2).

Table 9.2: Representation of the respirometry model as a Gujer matrix consisting of state variables to represent aerobic degradation of acetate $\mathrm{S_A}$ by biomass $\mathrm{X_B}$ consuming oxygen $\mathrm{S_O}$

process p_{i}	stoichiometry			reaction rate ρ_{i}
	$\mathrm{X_B}$	$\mathrm{S_A}$	$\mathrm{S_O}$	
Heterotrophic growth with $\mathrm{S_A}$ as substrate	1	$-\frac{1}{Y}$	$-\frac{1-Y}{Y}$	$\mu_{\max} \frac{\mathrm{S_A}}{K_{\mathrm{S}}+\mathrm{S_A}} \mathrm{X_B}$
Endogenous respiration	-1		-1	$b\mathrm{X_B}$
Aeration			1	$k_{\mathrm{La}}(S^0_{\mathrm{O}} - \mathrm{S_O})$
stochiometric parameters: Y	biomass $(\mathrm{M_{(COD)}l^{-1}})$	substrate $(\mathrm{M_{(COD)}l^{-1}})$	oxygen $(\mathrm{M_{(-COD)}l^{-1}})$	kinetic parameters: $\mu_{\max}, K_{\mathrm{S}}, b$ $k_{\mathrm{La}}, S^0_{\mathrm{O}}$

9.5 A Gujer matrix alternative for hydrology

Despite the differences with chemical reactions, where the matrix can be used to distinguish between the stochiometric and kinetic coefficients, the idea of combining the processes, state variables and fluxes in a single matrix is reusable with

respect to the current range of lumped hydrological modelling described in literature.

To translate the concept of the Gujer matrix to a hydrological point of view, we need to translate existing lumped hydrological models such as those created by the pyfuse environment into their respective components. From a functional process perspective, catchment dynamics include partition, storage, release, and transmission of water (Fenicia et al., 2011). These are represented by the usage of three generic building blocks:

1. Reservoir element: represents storage and release of water
2. Lag function element: represents the transmission and delay of fluxes
3. Junction element : represents the splitting, merging, and/or rescaling of fluxes

Different configurations of these building blocks can be constructed to represent the catchment characteristics. Furthermore, constitutive functions (e.g., relating fluxes to reservoir storage) and associated parameters need to be defined to construct new models. As such, these building blocks and constitutive functions need to be represented in the proposed matrix representation, inducing some adaptations to the original Gujer concept. A major difference with the Gujer matrix is the description of the transport terms by the matrix instead of the conversion functions (and related stoichiometry). The proposed matrix representation counterpart for lumped hydrological model structures is drafted in Table 9.3. For each part, the incorporation will be discussed.

9.5.1 Reservoir element

A reservoir element in lumped hydrological modelling is a representation of catchment scale processes related to storage and release of water. As such, this can be represented by mass balances, i.e. a set of ODEs (Equation 2.1), where each reservoir models the storage of water in function of time of a represented catchment entity. The incoming and outgoing fluxes are defined by either external forcing (e.g. rain, evapotranspiration), internal fluxes (e.g. percolation, drainage...) or outgoing fluxes (discharge). The response observed and used for evaluation is in the case of hydrological modelling mostly a discharge (flux), or any aggregation metric derived from it (cfr. section 3.3). As such, the original model definition of equation 2.1 can be translated to:

Table 9.3: Translation of the Gujer matrix concept towards standard matrix representation of lumped hydrological models consisting of state variable $S_1,\ldots,\ S_m$, processes $p_1,\ldots,\ p_n$, flux redistribution indicated by $\nu_{1,1},\ldots,\ \nu_{n,m}$ and constitutive functions describing the fluxes $f_1,\ldots,\ f_n$

process p_i	reservoir configuration				flow	constitutive functions
	S_1	S_2	...	S_m	q_{tot}	
p_1	$\upsilon_{1,1}$	$\upsilon_{1,2}$	$\cdots$	$\upsilon_{1,m}$		f_1
p_2	$\upsilon_{2,1}$	$\upsilon_{2,2}$	$\cdots$	$\upsilon_{2,m}$		f_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
p_i	$\upsilon_{i,1}$	$\upsilon_{i,2}$	$\cdots$	$\upsilon_{i,m}$	$q_i^* h_f(t)$	f_i
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
p_n	$\upsilon_{n,1}$	$\upsilon_{n,2}$	$\cdots$	$\upsilon_{n,m}$	q_n	f_n
lag functions $h_f(t)$	reservoir type	reservoir type	$\vdots$	reservoir type		parameters overview

$$\frac{\mathrm{d}\mathbf{S}(t)}{\mathrm{d}t} = f(\mathbf{S}(t), \mathbf{q}_{t,in}(t), \boldsymbol{\theta}, t) \tag{9.2}$$

$$\hat{\mathbf{q}}(t) = g(\mathbf{S}(t), \mathbf{q}_{t,in}(t), \boldsymbol{\theta}, t) \tag{9.3}$$

for n_s defined reservoir elements, with state variables $\mathbf{S}(t)$ representing storage, subject to external forcing $\mathbf{q}_{t,in}(t)$ and fluxes $\hat{\mathbf{q}}(t)$ that can be related to a measured variable.

The mass balances define the incoming and outgoing fluxes of the reservoir. Each mass balance is represented by a column in the matrix (reservoir configuration) and the processes p_i acting on the reservoir are listed in the rows of the matrix. The incoming and outgoing fluxes for each specific reservoir are listed as fluxes $\upsilon_{i,j}$, defined by the flux name and a positive or negative sign, representing respectively incoming or outgoing flow. In the last row, a full description of the reservoir type can be added to clarify the catchment function representation of the reservoir.

Each of the processes defining fluxes $\upsilon_{i,j}$ is defined by either a constitutive function (f_i) or an external forcing ($q_{t,in}$). The description is listed in the last column of

the matrix representation and can vary widely. As mentioned by Fenicia et al. (2011), these functions will form part of an extendable library with some of them frequently chosen. The concept is comparable to typical kinetic functions that are used in (bio)chemical reaction descriptions, with a - sometimes preconceived - preference of Monod kinetics.

In most cases, the observed flux $\hat{\mathbf{q}}(t)$ is the combination of outgoing fluxes coming from different reservoir elements (mostly the catchment outflow). This is represented by a separate column defining $\mathrm{q_{tot}}$, where all the contributing fluxes are listed and the sum of the individual fluxes provide a comparison with the measured catchment outflow. In the case of subflow comparison (Willems et al., 2014), individual fluxes can be linked to the subflows measured (or derived with filtering techniques).

9.5.2 Junction element

In contradiction to a typical Unit Hydrograph approach for routing application in lumped hydrological models which is a consecutive set of linear reservoirs, the representation of the entire set of catchment processes is mostly an interconnection of reservoirs in function of the catchment characteristics. Multiple reservoirs (and lag-functions) are connected with each other using junction functions (either joining or splitting). A typical example is the joining of fluxes coming out of reservoirs before entering yet another reservoir. These junction elements can also contain parameters to manipulate the junction.

The representation of junction elements is embedded in the reservoir configuration and they are part of the $\upsilon_{i,j}$ elements. Functions and parameters are written as a matrix element within the reservoir configuration. The rule is that the specific flux $\mathrm{q_i}$ used in that line is described by the constitutive function in the rightmost column of the matrix. As such, other elements (parameters and/or functions) can be used to represent splits or joints, next to lag functions discussed in the next section 9.5.3.

Direct joints of two reservoirs into a third reservoir are captured by the format itself, where two negative fluxes will appear (on different columns) and with a positive sign at the column of the receiving third reservoir. A split can be achieved on a single line, as illustrated in Table 9.5 to redistribute the saturation excess $\mathrm{q_q}$ (which itself is calculated by the constitutive function at the right hand side). Hence, $\upsilon_{3,2} = +(1-d)\mathrm{q_q^*}h_\mathrm{f}(t)$ and $\upsilon_{3,3} = +d\mathrm{q_q}$ to divide it amongst respectively

reservoirs S_f and S_s. Hence, the constitutive function in the rightmost column describes the q_q calculation.

9.5.3 Lag function element

Delays arising from channel routing are present in many model descriptions and thus a necessary element to properly represent the catchment behaviour. Traditionally, models like Hydrologiska Byråns Vattenbalansavdelning model (HBV) and FUSE make a distinction between the storage part of the model and the routing part of the model. In those cases, the retention of water from channel routing is represented by a sequence of linear reservoirs.

Actually, reservoir configuration for routing could be explicitly incorporated in the matrix representation by adding an extra column for each reservoir in the cascade, including more state variables. Representing each individual reservoir of the routing sequence as individual columns in the matrix would hinder the interpretation of the matrix representation. However, in most cases, these tanks in series are assumed to behave linearly. As such, the link with the unit hydrograph concept (Beven, 2012) will be used to represent the routing of water as a lag-function (in the case of linear operators) instead of adding individual reservoirs in the matrix.

In general, the lag-function is represented by a convolution operator (e.g. Gamma-function, Nash-cascade...) acting on a described flux by adding $^*h_f(t)$. Fenicia et al. (2011) advocate to make those lag functions applicable in all parts of the model structure to provide flexibility beyond the traditional storage-routing model structure distinction. Hence, such a convolution operator can be added at the following locations:

- Added to a flux q_i in the flow column of the matrix representation. As such, this represents the traditional case of a routing part of the model, where an outgoing flux is routed to the catchment outlet. For example, in Table 9.3, the total discharge is calculated as $q_{tot} = q_i \, ^*h_f(t) + q_n$.
- Added to an internal catchment flux q_i as part of the reservoir configuration. An outgoing flux q_i of a reservoir S_1 is subject to the convolution operator before it enters in another reservoir S_2. Hence, the incoming flow of S_2 is $q_i \, ^*h_f(t)$. This is also illustrated in Table 9.5.
- Hypothetically it can also be added to an incoming external forcing. However, this application would probably be rather rare.

In the specific case of a lag-function affecting joined fluxes as illustrated in Figure 9.1, the matrix representation does not provide a direct representation. However, due to the linear properties of the lag functions used and taking into account superposition of linear functions, this situation is similar to applying twice the lag function to each of the fluxes individually. The latter can easily be incorporated in the matrix representation.

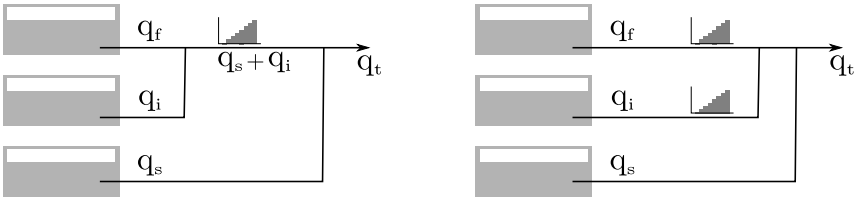


Figure 9.1: Situation example of a routing where the combination of two subflows q_f and q_i is affected by a lag function, wherafter the sum with a third flow q_s corresponds to the total outflow (left). This situation cannot be directly represented by the matrix representation. However, due to the linear characteristics of the lag functions used to represent routing, this is similar to the representation where both are affected indivually by a lag function (right) which can be easily incorporated in the matrix notation.

9.6 Application to existing model structures

In order to test and illustrate the usability of the matrix representation, some existing models will be converted into the proposed matrix format. First, two model structures used in Kavetski and Fenicia (2011) are converted to the matrix representation. These models are referred as model M1 and model M7 similar to the model names in Fenicia et al. (2011) and Kavetski and Fenicia (2011). Both M1 and M7 are already defined as a set of ODEs in the original publication and the matrix is provides a more dense representation.

Next, two models regularly used in both an operational and scientific setting, respectively PDM Moore (1985) and NAM Nielsen and Hansen (1973), will be handled. Current literature does not provide these model structures as a set of ODEs. Hence, their translation towards a set of ODEs is required before the matrix representation for these models can be defined.

9.6.1 Model M1 (Kavetski and Fenicia, 2011)

Model M1 is a minimalistic representation of a catchment by representing the entire catchment as a perfectly mixed reactor (Figure 9.2). Despite the limited usability of this structure in real applications, the representation provides an easy first example of the matrix representation concept.

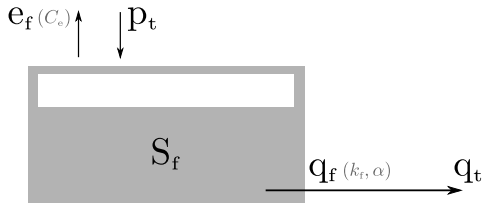


Figure 9.2: Model M1, acting as the extreme simplistic model representation, consists of a single non-linear reservoir S_f with three parameters. The outflow q_f is a power function of the storage, while the predicted evaporation e_f is proportional to the potential evaporation (Kavetski and Fenicia, 2011). Model parameters are added in gray font color.

The model describes three main processes: rain, evaporation and catchment outflow (left column of Table 9.4). The model exists of a single non-linear reservoir, represented in the reservoir configuration part of the table as a single mass balance:

$$\frac{dS_f}{dt} = p_t - e_f - q_f \tag{9.4}$$

with p_t represented by the incoming rainfall $p_{t,in}$ acting as an external forcing. Evaporation is proportional to the potential evaporation $e_{t,in}$, which is an external forcing as well. The storage S_f influences both the constitutive functions of evaporation and outflow. Evaporation is defined by parameter C_e and a smoothing function for near-zero storage values, governed by a smoothing parameter ω. Outflow q_f is described by a power function of the storage with parameters α and k_f. No lag-functions are used in model M1, the total flow $q_{tot} = q_f$.

9.6.2 Model M7 (Kavetski and Fenicia, 2011)

Model M7 consists of three reservoirs, eight parameters and one lag function. Hence, it resembles more complex model representations actually used for practical applications. The unsaturated reservoir S_u receives incoming rain $p_{t,in}$, evaporates

Table 9.4: Gujer matrix representation of the M1 lumped hydrological model structure presented in Kavetski and Fenicia (2011)

process	reservoir configuration S_f	flow q_{tot}	constitutive functions
rain	$+p_t$		$p_{t,in}$
evaporation	$-e_f$		$e_{t,in} C_e \left(1 - e^{-\frac{S_f}{\omega}}\right)$
outflow	$-q_f$	q_f	$k_f (S_f)^{\alpha}$
	fast flow		parameters C_e, k_f, α

water $-e_f$ and produces excess overflow water $-q_q$, which is divided amongst the other reservoirs S_f and S_s.

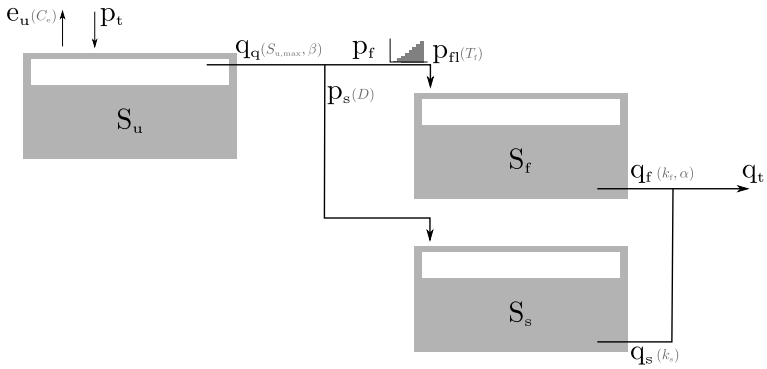


Figure 9.3: Model M7, a three reservoir lumped hydrological model with in total eight parameters (Kavetski and Fenicia, 2011). The excess water of the unsaturated reservoir S_u is distributed amongst a fast flow reservoir S_f and a groundwater reservoir S_s. Model parameters are added in gray font color.

A lag function affects the sub flux going to the fast flow reservoir and is explained in more detail in the left lower corner of the matrix representation. The fast flow reservoir acts as a non-linear routing function, whereas the slow flow reservoir represents the catchment groundwater by means of a single linear reservoir. The lower right corner of the matrix representation gives an overview of the parameters of the constitutive functions, the flux split and the lag function. Total flow is derived from summing up the fluxes listed in the flow column $q_{tot} = q_f + q_s$.

Table 9.5: Gujer matrix representation of the M7 lumped hydrological model structure presented in Kavetski and Fenicia (2011). The operator * denotes a convolution operator to incorporate lag functions in the model structure representation

process	reservoir configuration			flow	constitutive functions
	S_u	S_f	S_s	q_{tot}	
rain	$+p_t$				$p_{t,in}$
evapo-transpiration	$-e_u$				$e_{t,in} C_e \frac{(1+\omega)\frac{S_u}{S_{u,max}}}{\frac{S_u}{S_{u,max}}+\omega}$
saturation excess	$-q_q$	$+(1-d)\cdot q_q{}^* h_f(t)$	$+d q_q$		$p_{t,in}\left(\frac{S_u}{S_{u,max}}\right)^{\beta}$
fast routing		$-q_f$		q_f	$k_f(S_f)^{\alpha}$
slow routing			$-q_s$	q_s	$k_s S_s$
lag functions $h_f(t) =$ $\begin{cases} \frac{t}{T_f^2} & \text{if } t \le T_f \\ 0 & \text{if } t > T_f \end{cases}$	unsaturated	fast flow	slow flow		parameters $C_e, S_{u,max}, \beta, k_f, \alpha, k_s, d, T_f$

9.6.3 NAM model

Original NAM model

NAM is the abbreviation of the Danish Nedbør Afstrømnings Model, literally meaning rainfall-runoff model. Nielsen and Hansen (1973) describe the original model, developed at the Hydrological section of the Institute of Hydrodynamics and Hydaulics Engineering at the Technical University of Denmark. During the last decade, the model is maintained by DHI (Danish Hydraulic Institute) as a part of the MIKE software-suite. It is used within the operational water management at the Flanders Hydraulics Research, a division of the department of Mobility and Public Works of the Flemish government.

The NAM model is a rainfall-runoff model that operates by continuously accounting for the moisture content in different and mutually interrelated storages. These storages include: (1) snow storage (not included here), (2) surface storage U, (3)

lower or root zone storage L and (4) groundwater storage (S_3) (DHI, 2008). The model structure is shown in Figure 9.4. In the remainder of the section, the original naming U and L is used to make the parallel with the original model description and parameter names.

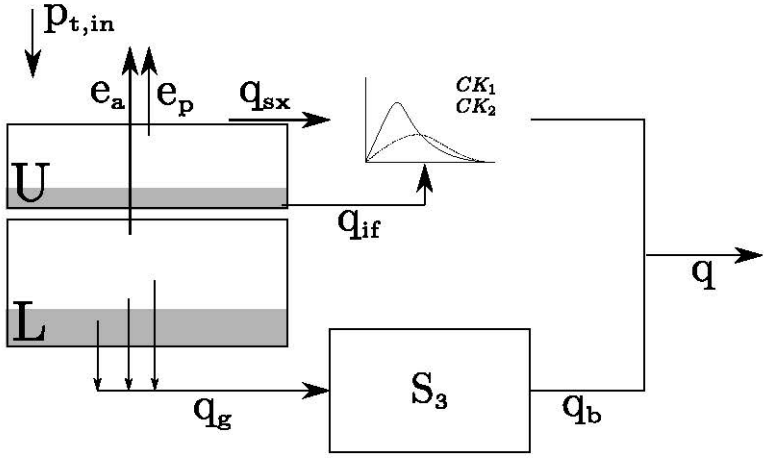


Figure 9.4: Overview of the original NAM model, illustrating the surface U storage reservoir and lower soil storage L reservoir representing the soil compartment, the overland routing and the base flow routing by reservoir S_3 (scheme redrafted from DHI (2008)).

Rainfall contributes to the surface storage when the temperature is above freezing point (freezing is neglected for this dissertation and not shown in figure). When the surface storage compartment is full, the remaining rainfall infiltrates towards the lower zone storages and contributes to the overland flow. Water is also extracted by (potential) evapotranspiration and interflow (hypodermic flow, i.e. horizontal flows in the unsaturated zone). The lower zone storage controls the different subflows, varying linearly with the relative soil moisture content of this lower zone storage. The different processes modelled by NAM are conceptualized by 9 empirical model parameters that need to be calibrated. A short description of each one of the model parameters is presented in Table 9.6.

The potential evapotranspiration, $e_{t,in}$, is a forcing variable. The evapotranspiration of the surface storage e_p occurs at a potential rate and is limited by the available water content ($e_p = e_{t,in} - U$). When the moisture content U is less than potential evapotranspiration $e_{t,in}$, the remaining fraction of evapotranspiration varies linearly with the lower storage water content (L/L_{max}) by:

Table 9.6: Overview of the NAM model parameters of the original model description (DHI, 2008)

parameter		**description**
$U_{\max}$	mm	Maximum water content in the surface storage
$L_{\max}$	mm	Maximum water content in the lower zone
CQ_{OF}		Overland flow runoff coefficient
T_{OF}		Threshold value for overland flow recharge
T_{IF}		Threshold value for interflow recharge
T_{G}		Threshold value for groundwater recharge
CK_{IF}	h	Time constant for interflow from the surface storage
$CK_{1,2}$	h	Time constant for overland flow and interflow routing
CK_{BF}	h	Time constant for base flow routing

$$\mathrm{e_a} = (\mathrm{e_{t,in}} - \mathrm{U}) \cdot \frac{\mathrm{L}}{L_{\max}} \tag{9.5}$$

Total evapotranspiration is modelled as the sum of $\mathrm{e_p}$ and $\mathrm{e_a}$. The interflow (hypodermic flow), $\mathrm{q_{if}}$, is assumed to be proportional to the surface storage U, and is given as

$$\mathrm{q_{if}} = \begin{cases} \frac{1}{CK_{\mathrm{IF}}} \frac{\frac{\mathrm{L}}{L_{\max}} - T_{\mathrm{IF}}}{1 - T_{\mathrm{IF}}} \mathrm{U} & \text{if} \frac{\mathrm{L}}{L_{\max}} > T_{\mathrm{IF}} \\ 0 & \text{if } \frac{\mathrm{L}}{L_{\max}} \leq T_{\mathrm{IF}} \end{cases} \tag{9.6}$$

When surface storage is full, excess rainfall $\mathrm{p_n}$ (effective rainfall after subtracting the interflow), will form overland flow, whereas the remainder is diverted as infiltration into the lower zone and groundwater storage. Overland flow, $\mathrm{q_{sx}}$, is assumed to be proportional to this saturation excess $\mathrm{p_n}$ and depends on the soil moisture content in the lower zone storage, given as

$$\mathrm{q_{sx}} = \begin{cases} CQ_{\mathrm{OF}} \frac{\frac{\mathrm{L}}{L_{\max}} - T_{\mathrm{OF}}}{1 - T_{\mathrm{OF}}} \mathrm{p_n} & \text{if } \frac{\mathrm{L}}{L_{\max}} > T_{\mathrm{OF}} \\ 0 & \text{if } \frac{\mathrm{L}}{L_{\max}} \leq T_{\mathrm{OF}} \end{cases} \tag{9.7}$$

The amount of water recharging the groundwater storage depends on the soil moisture content in the root zone. The groundwater storage will generate baseflow. The baseflow is assumed to be proportional to the amount of infiltrating water recharging the groundwater storage and depends on the soil moisture content in the lower zone storage. The groundwater recharge is given by

$$q_g = \begin{cases} (p_n - q_{sx}) \frac{\frac{L}{L_{max}} - T_G}{1 - T_G} & \text{if } \frac{L}{L_{max}} > T_G \\ 0 & \text{if } \frac{L}{L_{max}} \leq T_G \end{cases} \tag{9.8}$$

The routing of the inter flow uses two linear reservoirs in series with the time constants CK_1 and CK_2, usually assumed equal ($CK_{1,2}$). Overland routing is also based on two linear reservoirs, but with a variable time constant depending on an upper limit for linear routing (equation not given, analytical solution used). The base flow q_b routing is calculated as the outflow from one linear reservoir (S_3 in Figure 9.4) with time constant CK_{BF}. Total flow q is assessed by summing all different subflows.

The original NAM model uses an Operator Splitting (OS) approach in combination with an explicit fixed step solver to calculate the states and flows in function of time. Due to the closed source properties of the code implementation, further evaluation of the implementation is however limited.

ODE representation of NAM

When screening the general structure of the NAM model, the model can be separated by a storage part and a routing part of the model. The surface storage and lower storage are accounting for the storage part of the model, whereas the base flow compartment can be seen as part of the routing when the capillary flow is not taken into account (as is assumed regularly). Doing so, the routing model of the NAM model can be categorized as a set of linear reservoirs for the different subflows. Hereby, the further representation of the NAM model will only focus on the storage part.

As opposed to other conceptualizations, the water storage representation in the NAM model upper storage represents storage of water that is intercepted by vegetation, captured in surface depressions and storage in the uppermost layers (a few cm) of the soil. In Figure 9.4, the similarity with a soil moisture profile is made (DHI, 2008), where the upper soil storage represents the fraction above field capacity (free storage), filled when the tension storage is at capacity. Actually, the conceptualization is slightly different and the water movement from the upper storage to the lower storage can be interpreted as excess water of the upper storage that is diverted as infiltration towards the lower zone. The excess water that is not infiltrating will enter the streams as overland flow, for which it can be interpreted as infiltration excess as well. Similar to the original NAM, U and L are

used, where the U represents a surface storage transferring water to the tension storage L when at capacity.

The state equations for the NAM implementation are:

$$\begin{aligned} \frac{\mathrm{dU}}{\mathrm{d}t} &= \mathrm{p_{t,in}} - \mathrm{e_p} - \mathrm{q_{if}} - \mathrm{q_{sx}} - \mathrm{q_b} - \mathrm{q_{stof}} \\ \frac{\mathrm{dL}}{\mathrm{d}t} &= \mathrm{q_{stof}} - \mathrm{e_a} \end{aligned} \tag{9.9}$$

with the fluxes given in Table 9.7, where the different smoothing functions are added. The flux $\mathrm{q_{stof}}$ is added in addition to the original model description. This does not alters the conceptual idea of the NAM model, but is required to represent the overflow of water when the maximum storage capacity $U_{\max}$ is reached. The overflow is the amount of water flowing to the lower zone storage, which is similar to the original NAM conceptualisation. Whereas a maximal storage capacity $L_{\max}$ for the lower zone is defined as parameter as well, this is only used to calculate the fluxes while conceptualizing the storage itself as of unlimited size (no overflow of water).

Compared to the original constitutive functions defined in section 9.6.3, additional smoothing operators Φ are used as well. These operators also do not change the conceptual model, but are added to improve the handling of threshold-type behaviour, which can result in discontinuities in the response surface (Clark and Kavetski, 2010; Kavetski and Clark, 2010). Φ represents a logistic smoothing operators as used by Clark et al. (2008) (and included in Table 9.8). For a more in depth discussion in terms of implications and possible solutions, the reader is referred to Kavetski and Kuczera (2007).

The ODE representation differs from the original NAM version, giving rise to other flow calculations. Figure 9.5 compares the flow outcomes of both versions for a three year period, for both the calculated outflow and the different subflows. Figure 9.6 shows a comparison between the state variables in both implementations. The effect on the resulting outflow seems rather small, but the differences on the state variables and the individual subflows is larger. This is because the model is slightly different conceptualized to enable a representation in terms of differential equations rather than the operator splitting approach explained in DHI (2008). The similarity is still appropriate to refer as the NAM model.

Table 9.7: Overview of the NAM fluxes in the ODE representation

Flux	Flux equation
evapotranspiration [a]	$\mathrm{e_p} = \mathrm{e_{t,in}}(1 - e^{-\frac{\mathrm{U}}{\omega}})$ $\mathrm{e_a} = (\mathrm{e_{t,in}} - \mathrm{e_1})\frac{\mathrm{L}}{L_{\max}}$
overland flow [b]	$\mathrm{q_{sx}} = CQ_{\mathrm{OF}}\Phi\left(\frac{\mathrm{L}}{L_{\max}}, T_{\mathrm{OF}}, \omega\right) \cdot \left[\frac{\frac{\mathrm{L}}{L_{\max}} - T_{\mathrm{OF}}}{1 - T_{\mathrm{OF}}}\mathrm{U}\right] \Phi\left(\mathrm{U}, U^{\max}, \omega\right)$
inter flow [b]	$\mathrm{q_{if}} = CK_{\mathrm{IF}}\left[\frac{\frac{\mathrm{L}}{L_{\max}} - T_{\mathrm{IF}}}{1 - T_{\mathrm{IF}}}\mathrm{U}\right] \Phi\left(\frac{\mathrm{L}}{L_{\max}}, T_{\mathrm{IF}}, \omega\right)$
base flow [b]	$\mathrm{q_b} = \Phi\left(\frac{\mathrm{L}}{L_{\max}}, T_{\mathrm{G}}, \omega\right) \cdot \left[\frac{\frac{\mathrm{L}}{L_{\max}} - T_{\mathrm{G}}}{1 - T_{\mathrm{G}}}\mathrm{U}\right] \Phi(\mathrm{U}, U_{\max}, \omega)$
overflow flux [b]	$\mathrm{q_{stof}} = \mathrm{p_{t,in}}\Phi\left(\mathrm{U}, U_{\max}, \omega\right)$

[a] Smoothing constraint for min function as proposed by Kavetski and Kuczera (2007)

[b] Smoothing step discontinuity by logistic smoothing as proposed by Kavetski and Kuczera (2007)

Matrix representation of NAM

The matrix representation is given in Table 9.8. Evapotranspiration of the surface and tension storage is split into two separate processes. Both are called evapotranspiration, notwithstanding the different interpretation given to both. Since the water in the surface storage is considered to be freely available, this could be noted as evaporation instead of evapotranspiration. However, to remain similarity to the description in the model manual (DHI, 2008), the usage of evapotranspiration for both is preserved.

The constitutive functions for overland flow, inter flow and base flow are very similar and as such, part of the function $(g(T_{\mathrm{x}}))$ is summarized in the parameter section of the matrix representation to support readability. Functions for smoothing the differential equations are added in the short notation as well. In literature

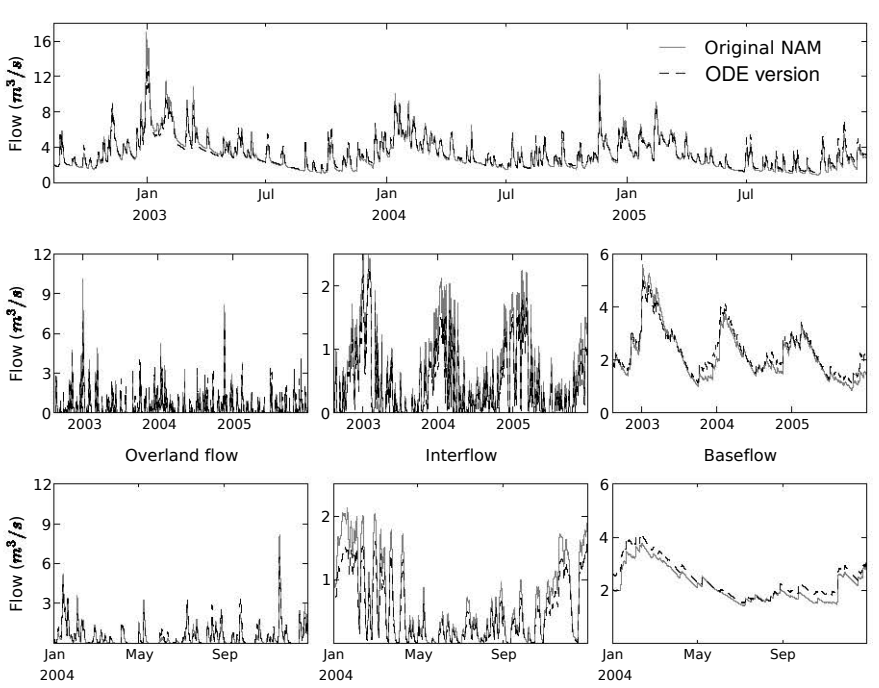


Figure 9.5: Comparison of the fluxes calculated by the original NAM implementation and the representation as ODEs. The combined outflow for a three year period (a), the three subflows for the same period (b) and a zoom on 2004 of the subflows (c) is presented.

papers, these operators and their smoothing parameters should be provided as well to ensure reproducibility of the model structure implementation.

The routing components of the model consistent of linear reservoirs and are added as lag functions of the different subflows, with the subscript n defining the number of tanks. In the case of the single reservoir base flow routing, the Gamma function reduces to the analytical solution of a single reservoir. To understand the similarity with Figure 9.4, it is important to understand that the lag-function that is combined with $\mathrm{q_g}$ corresponds to reservoir $\mathrm{S_3}$ (a linear reservoir) and the resulting flow derived from $\mathrm{q_g^* }h_{\gamma,1}(t)$ represents the baseflow $\mathrm{q_b}$. Total catchment outflow q is given by $\mathrm{q_{tot}} = \mathrm{q_{sx}^* }h_{\gamma,2}(t) + \mathrm{q_{if}^* }h_{\gamma,2}(t) + \mathrm{q_g^* }h_{\gamma,1}(t)$.

Table 9.8: Gujer matrix representation of the NAM lumped hydrological model structure. The naming 'et' is a short description of evapotranspiration. The operator * denotes a convolution operator to incorporate lag functions in the model structure representation. Function $g(T_\mathrm{x})$ is a help function to shorten notation, due to the similarity of the constitutive functions. Φ are smoothing functions to handle threshold behaviour as proposed by Kavetski and Kuczera (2007).

process	reservoir configuration U	reservoir configuration L	flow $\mathrm{q_{tot}}$	constitutive functions
rain	$+\mathrm{p_t}$			$\mathrm{p_{t,in}}$
surface et	$-\mathrm{e_p}$			$\mathrm{e_{t,in}}(1-e^{-\frac{\mathrm{U}}{\omega}})$
tension et		$-\mathrm{e_a}$		$(\mathrm{e_{t,in}}-\mathrm{p_1})\frac{\mathrm{L}}{L_\mathrm{max}}$
overland flow	$-\mathrm{q_{sx}}$		$\mathrm{q_{sx}^*}h_{\gamma,2}(t)$	$CQ_\mathrm{OF}\Phi\left(\frac{\mathrm{L}}{L_\mathrm{max}},T_\mathrm{OF},\omega\right)\cdot g(T_\mathrm{OF})\Phi\left(\mathrm{U},U_\mathrm{max},\omega\right)$
inter flow	$-\mathrm{q_{if}}$		$\mathrm{q_{if}^*}h_{\gamma,2}(t)$	$CK_\mathrm{IF}\Phi\left(\frac{\mathrm{L}}{L_\mathrm{max}},T_\mathrm{IF},\omega\right)g(T_\mathrm{IF})$
base flow	$-\mathrm{q_g}$		$\mathrm{q_g^*}h_{\gamma,1}(t)$	$\Phi\left(\frac{\mathrm{L}}{L_\mathrm{max}},T_\mathrm{G},\omega\right)\cdot g(T_\mathrm{G})\Phi(\mathrm{U},U_\mathrm{max},\omega)$
overflow flux	$-\mathrm{q_{stof}}$	$+\mathrm{q_{stof}}$		$\mathrm{p_{t,in}}\Phi\left(\mathrm{U},U_\mathrm{max},\omega\right)$
lag functions $h_{\gamma,n}(t)=\frac{1}{k\Gamma(n)}\left(\frac{t}{k}\right)^{n-1}e^{-\frac{t}{k}}$ with k equal to $CK_{1,2}$ or CK_BF	surface storage	tension storage		parameters $U_\mathrm{max},L_\mathrm{max},CQ_\mathrm{OF},CK_\mathrm{IF},T_\mathrm{OF},T_\mathrm{IF},T_\mathrm{G},CK_{1,2},CK_\mathrm{BF}$ and $g(T_\mathrm{x})=\left[\frac{\frac{\mathrm{L}}{L_\mathrm{max}}-T_\mathrm{x}}{1-T_\mathrm{x}}\mathrm{U}\right]$ and $\Phi(\mathrm{y},\mathrm{y_{max}},\omega)=\frac{1}{1+e^{\frac{\mathrm{y_{max}}-\mathrm{y}-\omega\epsilon}{\omega}}}$

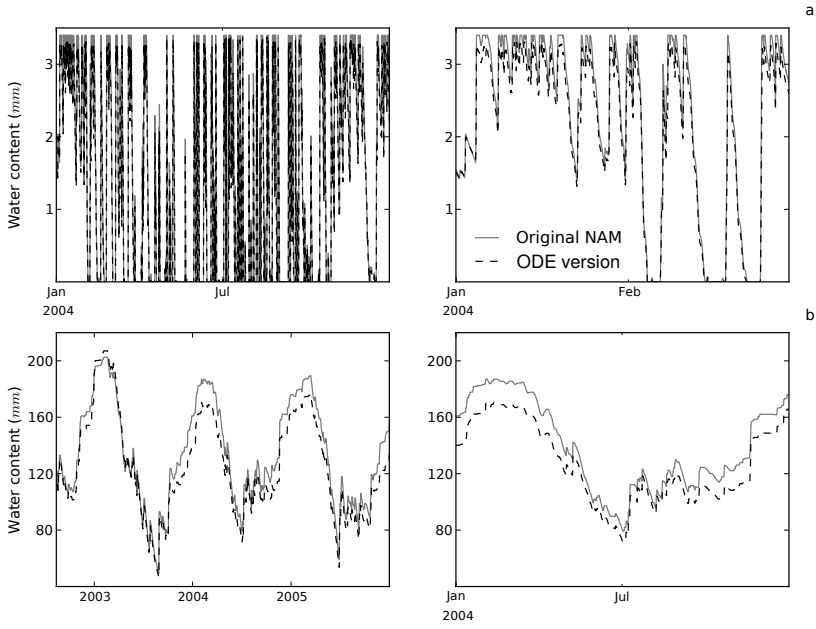


Figure 9.6: Comparison of the states calculated by the original NAM implementation and the representation as ODEs. The surface storage reservoir defined as U in the original NAM model (a) and the lower reservoir defined as L in the original NAM model (b) are presented. The periods shown are selected in function of the visual clarity, with for both (a) and (b) the right graph a zoom of the period shown in the left graph

9.6.4 PDM model

Original PDM model

The PDM is a lumped rainfall-runoff model which transforms rainfall and evaporation data into flow at the catchment outlet. Figure 9.7 shows the general layout of a PDM model that is commonly used in practice. The main model components are shortly discussed here and a more detailed description can be found in Moore (1985) and Moore (2007). It is used within the operational water management (flood forecasting) at the Flanders Environment Agency, part of the Environment, Nature and Energy policy domain of the Flemish government.

The model consists of three main components: (1) a probability distributed soil moisture storage component for separation of direct runoff and subsurface runoff,

(2) a surface storage component for transforming direct runoff to surface runoff (surface routing), (3) a groundwater storage which receives drainage water from the soil moisture storage component and contributes to baseflow (Moore, 2007). A description of the model parameters is presented in table 9.9.

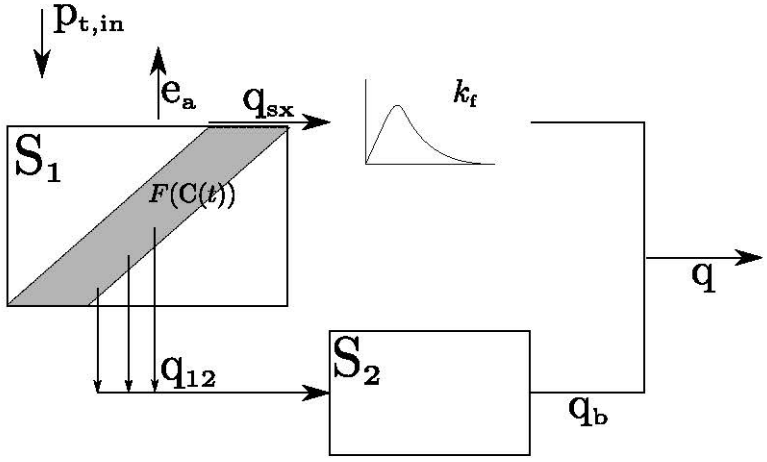


Figure 9.7: Overview of the PDM model structure illustrating the soil storage representation S_1, the routing of the overland flow with 2 linear reservoirs and the base flow reservoir, named S_3 in the original PDM description, but also referred to as S_2 to comply with the pyfuse layout (redrafted from Moore (2007))

The soil moisture storage component, defined by the probability distributions, represent different locations in the catchment, which also have different storage capacities. During any rain event, reservoirs with the smallest storage capacity will be filled first and will start to produce rapid runoff first. The area of the catchment that produces fast runoff is calculated from the proportion of the catchment with filled reservoirs $A_c(t)$. As such, the probability-distributed soil moisture storage component is used to separate direct runoff and subsurface runoff. Hence, the instantaneous direct runoff rate q_{sx} per unit area is defined by the product of the net rainfall rate ($p_{t,in} - e_a$) and the proportion of the basin generating runoff, defined by a distribution function $F(C(t))$ (see also equation 9.15).

A Pareto or truncated Pareto distribution function is mostly invoked for practical applications, although the PDM model offers a wide range of possible distributions (Moore, 2007). In this study, the following Pareto distribution function $F(C)$ and probability density function $f(C)$ are used to describe the critical capacity C below

which reservoirs are full at some time t:

$$F(\mathrm{C}(t)) = 1 - \left(1 - \frac{\mathrm{C}}{C_{\max}}\right)_{\mathrm{p}}^{b} \qquad 0 \leq \mathrm{C} \leq C_{\max} \tag{9.10}$$

$$f(\mathrm{C}(t)) = \frac{b_{\mathrm{p}}}{C_{\max}}\left(1 - \frac{\mathrm{C}}{C_{\max}}\right)^{b_{\mathrm{p}}-1} \qquad 0 \leq \mathrm{C} \leq C_{\max} \tag{9.11}$$

where $C_{\max}$ the maximum storage capacity in the basin and where parameter b controls the degree of spatial variability of storage capacity over the catchment. For the chosen Pareto distribution for storage capacity, the following unique relation between the storage over the basin as a whole $\mathrm{S}_1(t)$ and the critical capacity $\mathrm{C}(t)$ exists:

$$\mathrm{S}_1(t) = S_{1,\max}\left(1 - \left(1 - \frac{\mathrm{C}(t)}{C_{\max}}\right)^{b_{\mathrm{p}}+1}\right) \tag{9.12}$$

and the total available storage $S_{1,\max}$ can be derived from parameter $C_{\max}$ by $S_{1,\max} = \frac{C_{\max}}{b_{\mathrm{p}}+1}$.

The ratio between actual (e_{a}) and potential evapotranspiration ($\mathrm{e}_{\mathrm{t,in}}$) is defined as

$$\frac{\mathrm{e}_{\mathrm{a}}}{\mathrm{e}_{\mathrm{t,in}}} = 1 - \left(\frac{(S_{1,\max} - \mathrm{S}_1(t))}{S_{1,\max}}\right)^{b_{\mathrm{e}}} \tag{9.13}$$

and mostly depends linearly ($b_{\mathrm{e}} = 1$) or quadratically ($b_{\mathrm{e}} = 2$) on the soil moisture deficit, $(S_{1,\max} - \mathrm{S}_1(t))$.

Loss towards the groundwater as recharge is defined by the assumption that the rate of drainage, q_{12}, is linearly dependent on the basin soil moisture content:

$$\mathrm{q}_{12} = \frac{1}{k_{\mathrm{g}}}\left(\mathrm{S}_1(t) - S_\tau\right)^{b_{\mathrm{g}}} \tag{9.14}$$

where k_{g} is the drainage time constant and b_{g} the exponent of the recharge function, in this dissertation set to 1. S_τ is the threshold storage below which there is no drainage and the water is immobilised by the soil tension. Again, other drainage options are discussed in Moore (2007), but focus is here on the translation of a single chosen model structure.

In the original description of Moore (2007), both surface and base flow routing can be modelled by either non-linear storage reservoirs or a cascade of two linear

reservoirs. Here, a single (commonly applied) option is further used. The routing by the surface storage is represented by a cascade of two linear reservoirs, with equally assumed time constants k_{f}. Subsurface flow is routed by the groundwater storage by a non-linear storage routing function. In this case, baseflow is calculated by $\mathrm{q_b} = k_{\mathrm{b}} \left(\mathrm{S}_2(t)\right)^3$. By summing the surface runoff and base flow, the total discharge at the catchment outlet is calculated at every time step of the simulation. Notice that S_2 is used here, which does not correspond to the original model description of Moore (2007), referred to this reservoir as S_2.

ODE representation of PDM model

In the original PDM model (Moore, 2007), different distributions types are included to represent the probability-distributed storage model component. Nevertheless, in most applications a Pareto distribution is used as explained in section 9.6.4, which is similar to the VIC/ARNO model used by Clark et al. (2008) as well. Moore (2007) defines the critical capacity below which all storages are full at some time t as C and the contributing area A^* at time t for a basin of area A is:

$$A_{\mathrm{c}}(t) = \frac{A^*}{A} = F\left(\mathrm{C}(t)\right) \tag{9.15}$$

with the function $F(\mathrm{C})$ the distribution function of the storage capacity. The corresponding runoff $\mathrm{q_{sx}}$ is then defined by the fraction of rainfall as defined by

$$\mathrm{q_{sx}} = A_{\mathrm{c}} \mathrm{p_{t,in}} \tag{9.16}$$

The critical capacity C for the Pareto distribution is defined by:

$$\mathrm{C}(t) = C_{\max} \left(1 - \left(1 - \frac{\mathrm{S}_1(t)}{S_{1,\max}}\right)^{\frac{1}{b_{\mathrm{p}}+1}}\right) \tag{9.17}$$

If we combine equation 9.17 and equation 9.10, the contributing area $A_{\mathrm{c}}(t)$ is defined as

$$A_{\mathrm{c}}(t) = 1 - \left(1 - \frac{\mathrm{S}_1(\mathrm{t})}{S_{\max}}\right)^{\frac{b_{\mathrm{p}}}{b_{\mathrm{p}}+1}} \tag{9.18}$$

Table 9.9: Overview of the PDM model parameters

Parameter		Description
$C_{\max}$	mm	Maximum store capacity
b_{p}		Exponent of Pareto distribution controlling spatial variability of store capacity
b_{e}		Exponent in actual evaporation function
b_{g}		Exponent of recharge function
k_{g}	h $\mathrm{mm}^{(b_g-1)}$	Groundwater recharge time constant
k_{b}	h mm^2	base flow time constant
k_{f}	h	Time constants of cascade of two linear reservoirs
S_τ	h	Soil tension storage capacity

The non-linear baseflow reservoir of the original PDM model can be simulated by using a non-linear reservoir representing the lower layer storage with the baseflow exponent parameter $n = 3$.

The drainage is described by the flux q_{12} given in Table 9.10. The S_τ parameter defines the soil at field capacity, making the $\mathrm{S} - S_{\mathrm{tau}}$ conceptually identical to a free tension storage.

As such, we can combine these flux equations (an overview is provided in Table 9.10) in the following set of mass balances:

$$\begin{aligned} \frac{\mathrm{dS}_1}{\mathrm{d}t} &= \mathrm{p}_{\mathrm{t,in}} - \mathrm{e}_{\mathrm{a}} - \mathrm{q}_{\mathrm{sx}} - \mathrm{q}_{12} - \mathrm{q}_{\mathrm{ufof}} \\ \frac{\mathrm{dS}_2}{\mathrm{d}t} &= \mathrm{q}_{12} - \mathrm{q}_{\mathrm{b}} \end{aligned} \tag{9.19}$$

Similar to the translation for the NAM model, an additional overflow flux $\mathrm{q}_{\mathrm{ufof}}$ is defined when maximum capacity is reached. However, in the case of PDM, the overflow of the storage represents additional surface runoff. Furthermore, smoothing operators are added as well here to improve the handling of threshold behaviour (Kavetski and Kuczera, 2007).

Figure 9.8 compares the flow outcomes of the original PDM version as described by Moore (2007) and the ODE representation for a three year period, for both the calculated discharge and both subflows. Figure 9.9 focuses on the state variables. The differences of the modelled flow are smaller than the differences of the NAM model ODE representation.

Table 9.10: Overview of the PDM fluxes in the framework version

Flux	Flux equation
evapotranspiration	$\mathrm{e_a} = \mathrm{e_{t,in}} \left(1 - \left(1 - \frac{\mathrm{S_1}}{S_{1,\max}}\right)^{b_\mathrm{e}}\right)$
overland flow [a]	$\mathrm{q_{sx}} = \mathrm{p_{t,in}} A_\mathrm{c}$
percolation [b]	$\mathrm{q_{12}} = \frac{1}{k_\mathrm{g}} (\mathrm{S_1} - S_\tau)^{b_\mathrm{g}}\, \Phi(\mathrm{S_1}, S_\tau, \omega)$
base flow	$\mathrm{q_b} = k_\mathrm{b} (\mathrm{S_2})^3$
overflow flux [b]	$\mathrm{q_{ufof}} = (\mathrm{p_{t,in}} - \mathrm{q_{sx}}) \Phi(\mathrm{S_1}, S_{1,\max}, \omega)$

[a] Probability soil moisture store based saturated area A_c given in equation 9.18

[b] Smoothing step discontinuity by logistic smoothing as proposed by Kavetski and Kuczera (2007)

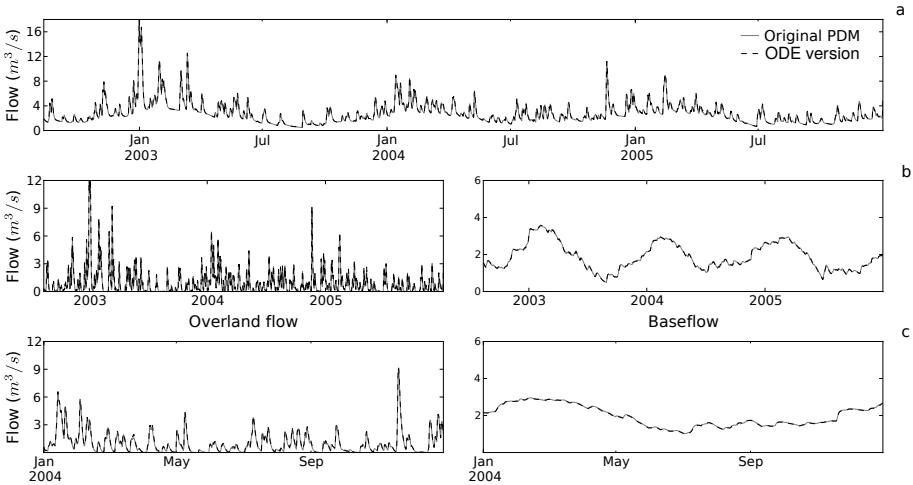


Figure 9.8: Comparison of the fluxes calculated by the original PDM implementation and the representation as ODEs. The combined outflow for a three year period (a), the two subflows for the same period (b) and a zoom on 2004 of the subflows (c) is presented.

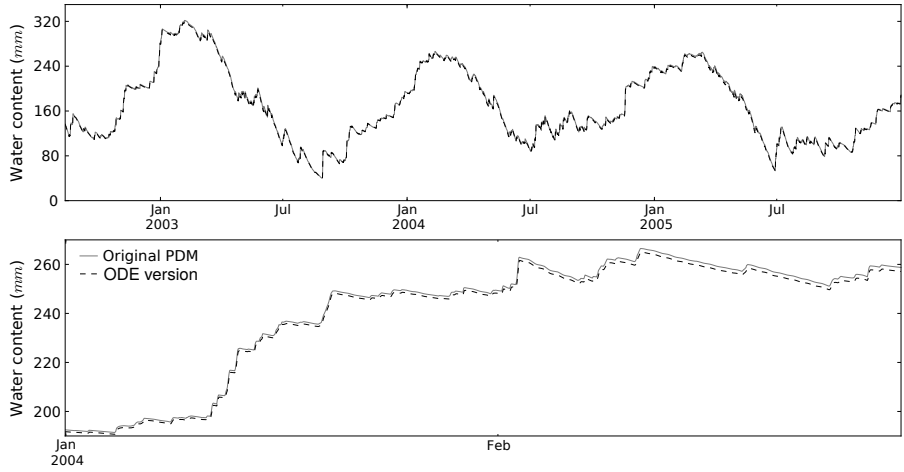


Figure 9.9: Comparison of the state of the soil storage S_1 calculated by the original PDM implementation and the representation as ODEs.

Matrix representation of PDM

Table 9.11 represents the structure in the matrix representation. The storage part is represented by reservoir S_1, where the Pareto function is used as a constitutive function for overland flow q_{sx}. As such, the flexibility of the PDM model considering the probability distribution function, is translated in the chosen constitutive functions for overland flow.

A second reservoir (i.e. mass balance) is added to represent the groundwater flow, modelled as a non-linear reservoir with a power function with parameters α and k_b. The overland flow routing is not added to the matrix representation as a separate column, since it is modelled by a cascade of two linear reservoirs, which is nothing more than the lag function $h_\gamma(t)$, represented by the function in the lower left corner. In general, it is advisable to include these linear reservoir sequences as lag functions instead of extra columns (i.e. reservoirs). By doing so, the matrix representation is more dense, but it also provides clarity in what is solved numerically (each column) and what is solved analytically (lag functions as analytical solution or approximated). By adding the function to the overland flow q_{sx}, the total outflow of the catchment is derived by $q_{tot} = q_{sx}{}^* h_\gamma(t) + q_{ufof}{}^* h_\gamma(t) + q_b$.

Table 9.11: Gujer matrix representation of the PDM lumped hydrological model structure. The operator * denotes a convolution operator to incorporate lag functions in the model structure representation. Φ are smoothing functions to handle threshold behaviour as proposed by Kavetski and Kuczera (2007).

process	reservoir configuration S_1	S_2	flow q_{tot}	constitutive functions
rain	$+p_t$			$p_{t,in}$
evapo-transpiration	$-e_1$			$e_{t,in}\left(1-\left(1-\frac{S_1}{S_{1,max}}\right)^{b_e}\right)$
percolation	$-q_{12}$	$+q_{12}$		$\frac{1}{k_g}(S_1 - S_\tau)^{b_g} \cdot \Phi(S_1, S_\tau, \omega)$
overland flow	$-q_{sx}$		$q_{sx}{}^* h_\gamma(t)$	$p_{t,in}\left(1-\left(1-\frac{S_1}{S_{1,max}}\right)^{\frac{b_p}{b_p+1}}\right)$
base flow		$-q_b$	q_b	$k_b S_2^3$
overflow flux	$-q_{ufof}$		$q_{ufof}{}^* h_\gamma(t)$	$(p_{t,in} - q_{sx}) \cdot \Phi(S_1, S_{1,max}, \omega)$
lag functions $h_\gamma(t) = \frac{1}{k_f \Gamma(2)}\left(\frac{t}{k_f}\right) e^{-\frac{t}{k_f}}$	unsaturated	slow flow		parameters $S_{1,max}, b_p, b_e, b_g, k_g, k_b, k_f, S_\tau$ and $\Phi(y, y_{max}, \omega) = \frac{1}{1+e^{\frac{y_{max}-y-\omega\epsilon}{\omega}}}$

9.7 Discussion

The matrix representation in this chapter provides an overview of a model structure configuration in a dense format (cfr. the combination of tables and scheme in Kavetski and Fenicia (2011) and Clark and Kavetski (2010)). Mass balance equations can easily be derived from the matrix by writing down each column in the reservoir configuration. Moreover, the risk of missing an element in the model description is decreased which supports the transparency and reproducibility of the work. Still, complete reproducibility is only provided by making the source code itself available, since the usage of the same numerical solver within different development tools can still provide differences in the results (Seppelt and Richter, 2005).

A such, the matrix representation can be used in any hydrological modelling paper to specify the specific modelling decisions. Moreover, a specific representation of some of the currently well-known models (as was done for the PDM or NAM model in this chapter) could be agreed on as reference models and get a specific code, similar to what the wastewater treatment community did with the ASM family to model wastewater treatment plants (Henze et al., 1983; Gujer and Larsen, 1995). This would for example lead to a clearly defined PDM_{P} model to define the Pareto distribution version of the PDM model. This can pave the way for standardisation and benchmarking in hydrological modelling, in order to systematically evaluate competing alternatives and prioritize model development needs (Clark et al., 2015a).

This chapter provided the general representation and further tests should be done to evaluate the usefulness for practical applications. At the same time, the benefits of the representation could already by exploited. By translating existing lumped hydrological models in a systems dynamics representation, a manifold of modelling and simulation platforms can be used, such as the pyideas Python Package 2 developed by Van Daele et al. (2015c). Moreover, users of the programming language R would be able to solve the models with deSolve (Soetaert and Petzoldt, 2010b), taking benefit of the compatible available modelling techniques for sensitivity and uncertainty analysis (Soetaert and Petzoldt, 2010a).

Eventually, automatic converters and Gujer matrix editors, as they are part of existing software such as WEST (Claeys, 2008) and Aquasim (Reichert, 1994) can be developed for lumped hydrological model building. Moreover, it also provides a solution for existing model software to communicate about their specific model implementation without the need of sharing all of their source code and this in an

elegant and complete way, supporting a closed source business model as it is still frequently seen in environmental modelling.

The method is also relevant in a spatially explicit approach of (hydrological) modelling. Mass balances acting on a single cell (local entity), where *cell* can be typical grid cells, Hydrological Response Units (HRUs) (Olivera et al., 2006) or Representative Elementary Watersheds (REWs) (Reggiani et al., 1998, 1999) can be written down by the matrix representation presented here. An extra representation would be necessary to represent the spatial processes (spatial configuration). In essence, this is a set of PDEs in which fluxes are represented by constitutive functions as well.

9.8 Conclusion

In chapter 2 the lack of flexibility in the model development process is identified as a bottleneck for an improved model based approach. This issue is specifically apparent in the case of so-called lumped hydrological models, a class of models frequently used and studied in hydrological modelling. Model structures are provided as monolithic implementations with limited flexibility and unclear separation between the mathematical and computational model.

This chapter proposes a matrix representation for lumped hydrological model structures to overcome these issues. By treating these model structures as a set of ODEs, flexibility on the (finest) process level is accomplished and variations on individual model component combinations made possible. Moreover, the definition as a set of ODEs supports a separation between the mathematical and computational model. Finally, the matrix representation provides a generic representation of the equations in the mathematical model to make the model configuration more transparent without depending on the implementation itself.

By sharing the matrix in combination with the numerical scheme used to solve the equations, the elements are available to accurately reproduce any lumped hydrological model structure that can be translated as a set of ODEs which supports an improved practice of model structure handling and representation as sought-after and defined as objective of this dissertation.

CHAPTER 10

Dynamic identifiability analysis based model structure evaluation considering rating curve uncertainty

Redrafted from

Van Hoey, S., Nopens, I., van der Kwast, J., and Seuntjens, P. (2015b). Dynamic identifiability analysis-based model structure evaluation considering rating curve uncertainty. *Journal of Hydrologic Engineering*, 20(5):1–17

10.1 Introduction

Water managers and related decision makers use lumped hydrological models for a variety of applications, ranging from forecasting models, for catchment characterisation and incorporating them in integrated applications. The ability of such a model to reproduce observations determines the credibility of the predictions provided by the model. However, uncertainty in data, model parameters and model structure hampers this evaluation. The aim of this chapter is to provide more insight in model structural failures by combining the components and elements explained and implemented in previous chapters.

Parameter identifiability enables the identification of model deficiencies (chapter 2). Different methodologies for sensitivity and identifiability analysis are implemented in the pystran Python Package 4, where DYNIA is of particular interest

due to the temporal analysis of parameter identifiability. By evaluating the model performance in function of time, periods for which the model is failing are identified while the parameter influence identifies which model component is the most important during these periods (Reusser and Zehe, 2011). This concept of temporal parameter identification to identify and analyse deficits in model structure has been introduced by Beck (1986). Parameter values are considered to have a fixed value within a model structure. When variation of the parameter as a function of time would be needed to improve the model behaviour, this can actually indicate a model deficiency.

The DYNIA technique will be applied to two model instances of the pyfuse modelling environment presented in chapter 9. As such, the architectural implementation of both models is the same and these models can be compared on their structural properties itself. The latter requirement would not be satisfied when comparing model structures from different model software environments.

The DYNIA approach fits in the metric oriented approach (section 3.2.2). A proper performance metric needs to be defined to evaluate the model performance in function of time. Considering the limitations of discharge measurements in natural rivers being dependent on a correct stage-discharge relation (rating curve), the uncertainty should be taken into account in the model evaluation, i.e. translated towards the performance metric. The limits of acceptability (section 3.4.3) anticipates for this in the performance metric construction.

This chapter allows the parameter values to change as function of time in order to detect model structure failures. In order to take into account the limitations provided by uncertain data, the data uncertainty is taken into account in the performance metric construction. The aim is to check for model structural deficiencies by using a dynamic evaluation of the parameter values, while using the uncertain values of the measured discharge instead of the deterministic values normally reported and applied.

First of all, the issue of rating curve uncertainty is shortly introduced and previous applications of temporal parameter identification for model structure evaluation are discussed. Then, the strategy applied in this chapter (using components introduced in earlier chapters) is explained. Further, the individual steps are discussed in more detail in the materials and methods section. The outcomes and a discussion on the applied strategy are closing the chapter.

10.2 Rating curve uncertainty

The inherent uncertainty in flow measurement restricts the ability to discriminate among competing hydrological model structures (Clark et al., 2011b). Taking into account the uncertainty on the rating curve in the model evaluation is thereby worthwhile investigating.

Measuring discharges in natural rivers is not straightforward, due to the heterogeneity of the river bed and river banks. However, the measurement of the water level itself is more obvious to do. In order to relate the (constantly) measured water levels with the effective discharges in the river, a relation is set up between water level and discharge, which is called a rating curve.

With regard to rating curve uncertainty, Di Baldassarre and Montanari (2009) distinguish (1) errors of the stage-discharge relation induced by interpolating and extrapolating of river discharge observations, (2) the presence of unsteady flow and (3) the seasonal variation of the roughness, with increasing errors when discharges increase. To determine the observational error from rating curve interpolation and extrapolation, Blazkova and Beven (2009) and Westerberg et al. (2011a) use a fuzzy regression method introduced by Hojati et al. (2005). Pappenberger et al. (2006) use a two-dimensional fuzzy membership function to evaluate the parameter combinations for the rating curve functions resulting in likelihood measures to compute uncertainty bounds in prediction. Krueger et al. (2010) and McMillan et al. (2010) further extended this concept by fitting the rating curve towards a subset of data points and checking consistency of the fit with the remaining points.

The incorporation of the uncertainty of the rating curve in model evaluation has been described in literature and most approaches use a time step based method. Beven (2006) proposed the extended GLUE approach as a way to partly avoid the subjectivity of the GLUE uncertainty analysis by translating the uncertainty of the discharge observations in *limits of acceptability* (Blazkova and Beven, 2009; Westerberg et al., 2011b; Krueger et al., 2010; Liu et al., 2009a). The limits of acceptability approach (section 3.2.2) directly fits within the metric oriented approach as a method to construct a performance metric, which can be used by a wide range of methods and it not restricted to GLUE applications only. The latter provides information about the effect on variability of the model output.

Beven (2008b) proposes fuzzy weighting functions (in most cases triangular) to assign time step based weights according to the level of performance. The time step based weights can be aggregated to a model performance metric. McMillan et al.

(2010) derive the complete PDF of the measured flow to form a likelihood function used in Bayesian inference parameter search. This results in higher parameter uncertainty and hence wider uncertainty bounds for flow predictions compared to the use of a deterministic rating curve based inference scheme.

10.3 Time variant parameter identifiability for model structure evaluation

Temporal analysis to evaluate the information content of data and to extract signature information is a valuable procedure to identify potential model deficits, already proposed by Beck and Young (1976) and Beck (1986). Traditionally, this was done for discrete models and by applying an extended Kalman Filter approach for recursive parameter estimation. More recently, de Vos et al. (2010) use temporal clustering to identify periods of hydrological similarity. Reusser and Zehe (2011) propose an approach to understand model structural deficits based on a combination of the type of model errors with parameter influence and model component dominance. Reichert and Mieleitner (2009) combine the estimation of time dependent model parameters with the degree of bias reduction to identify model deficiency.

Several scientists proposed the use of time-variant and stochastic parameters based on observations of variations in optimal parameter sets and of relations between the model states and the optimal parameter set (Beck and Young, 1976; Cullmann and Wriedt, 2008; Lin and Beck, 2007; Reichert and Mieleitner, 2009; Kuczera et al., 2006; Tomassini et al., 2009). This proposal can be linked to the Data-Based Mechanistic approach (DBM) that uses state-dependent parameters to identify non-linear systems (Young et al., 2001). The main argument for introducing stochastic parameter values is the inherent stochasticity of conceptual models due to spatial and temporal averaging (Kuczera et al., 2006). Next to this, Cullmann and Wriedt (2008) argue that state-dependent parameter changes should be incorporated in the formulation of process based models intended for long term simulations, hereby adapting to different environmental conditions. Muleta (2012) reports improved calibration and validation results when applying a season-based calibration approach. However, when using lumped hydrological models, the general assumption remains that model parameters are constant in time, given that catchment characteristics do not change within the time frame for which the model is developed. If parameter optima change in time, then the inference is that there

is a missing aspect in the model formulation and thus a model structural error (Abebe et al., 2010).

As mentioned, the idea of allowing parameters to vary in time to gain information about potential model structural improvements goes back to Beck and Young (1976) and the potential of learning from the behaviour of time-dependent parameters is higher than from corrections in model states (Reichert and Mieleitner, 2009). As such, the use of time-dependent parameters and identifiability evaluation as done by the DYNIA approach is a key strategy for model structure evaluation. The DYNIA approach improves the amount of information that can be obtained from the observed time series through the use of a moving window. Cullmann and Wriedt (2008) compared the optimised parameter set derived with the Gauss Marquardt Levenberg (GML) algorithm on event basis with the identifiable regions of the DYNIA approach and concluded that in most cases both coincide. Furthermore, by reorganizing the data according to the state variable (i.e. flow) instead of using the time series as such, a relation between the optimal parameter value and the observed flow is revealed. This leads to the suggestion of using state-dependent (transient) model parameters for models in operational conditions. Wriedt and Rode (2006) conclude the same when they observed a shift in the confidence range of a parameter that controls the inter flow volume at increased discharge. They also evaluated the evolution of the parameter identification range with growing window size and concluded that for most parameters a constant uncertainty range was obtained after one or two years of simulation. Lee et al. (2004) compare two model structures, with one of them a probability based model structure. Parameters were either non-identifiable over the entire time series or exhibited time-dependency in their optimal values. Seasonal variations of the optimum parameter values were consistent and much clearer than the variations between dry and wet years. They suggest improved model structures based on the correlation between the shifts in the posterior distributions of the parameters and the soil moisture storage dynamics. However, these adaptations did not result in a significant improvement in terms of representing the outflow hydrograph. Tripp and Niemann (2008) use the DYNIA approach to compare a PDM with a more physically based soil moisture representation model. They noticed structural errors in both models and also argued that the identifiability of a model parameter is not a sufficient reason to confirm the assumptions underlying the parameter occurrence. Indeed, the parameter that seemed the most stable and identifiable in a short period of time appeared to vary in time when evaluating a long time period. Nevertheless, the latter is actually just a consequence of the nature of these models, making them only suitable in a limited space and time frame. Abebe et al. (2010) apply the DYNIA approach on the HBV model and

retrieved for three out of five analysed parameters clearly defined periods with high information content in order to identify the parameter values. The relation between model parameters and soil moisture state was also highlighted.

10.4 Model structure evaluation strategy

In this chapter, we apply a combination of existing methodologies for model evaluation implemented in chapter 3 (Figure 10.1) and we apply this approach to evaluate and compare the NAM and PDM lumped hydrological model structures implemented in the pyfuse python package as explained in chapter 3. Different elements for improving the model evaluation and identification are considered and discussed. As shown in Figure 10.1, the approach of combining existing methodologies to enhance the model structure evaluation consists of:

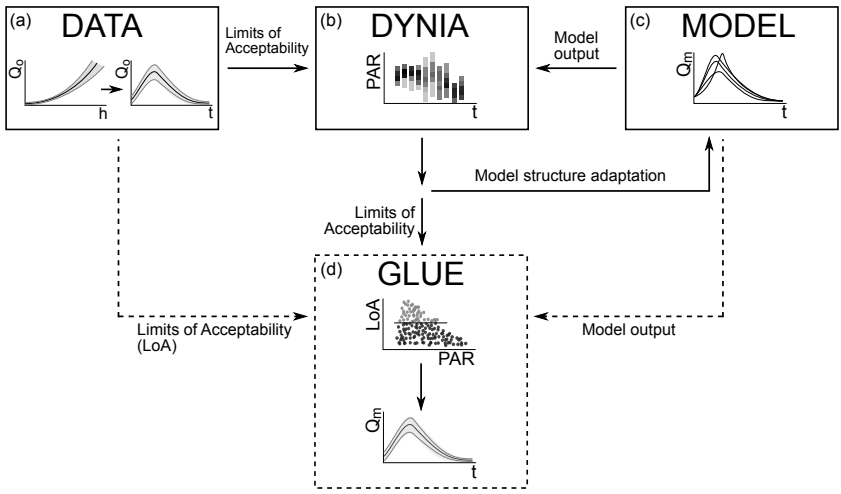


Figure 10.1: Schematic overview of the chapter. The DYNIA approach evaluates the model structure and parameter identifiability (b) based on limits of acceptability that are derived from the uncertainty in the rating curve (a) and on a Monte Carlo set of model runs (c). Subsequently, the prediction uncertainty is assessed with the GLUE approach (d).

- Figure 10.1a: Take data uncertainty into account of discharge observations. Since model performance metrics are based on the comparison of modelled and observed time series, they are very dependent on the reliability of the flow measurements used. The inherent uncertainty in the observed flows

often restricts the ability to discriminate among competing model structures (Clark et al., 2011a). Taking into account uncertainty on the rating curve in the model evaluation is thus essential, but usually not done in model evaluations.

- Figure 10.1b: A two-step application of the DYNIA approach is proposed and represents the central part of the methodology. By applying the DYNIA approach on a subset of selected simulations, it highlights the compensation of parameter values conditioned by the overall performance.
- Figure 10.1c: Incorporating two lumped hydrological model structures with different structural characteristics. The PDM model (Moore, 2007) uses a probability distribution to conceptualise the spatial differences in water storage capacity, whereas the NAM model (Nielsen and Hansen, 1973) assumes a single reservoir. Furthermore, different routing and groundwater configurations are used.
- Figure 10.1d: The lack of identifiability is further assessed by the GLUE approach by accepting all parameter sets and model structures that are behavioural according to the proposed limits of acceptability (Beven, 2006). The selected behavioural model simulations are used to compare the prediction uncertainty of both models under the defined acceptance limits. The results should be interpreted relatively in between both models to evaluate the effect of the identified model deficiencies on the prediction uncertainty.

As such, the workflow applied here attempts to provide maximal information about the malfunctioning of the models. By making the origin of malfunctioning more transparent, the modeller is less vulnerable to making type I and type II errors and gets more insight in the background of the prediction uncertainty.

The components of the method are laid out according to Fig. 10.1 both in section 10.5, Materials and Methods, as well as in section 10.6, Results. The latter section also contains the direct outcome of the model analysis. The reasoning about the structural deficiencies of the models used in the illustrating case together with the advantages and shortcomings of the combined approach are discussed in section 10.7.

10.5 Materials and Methods

10.5.1 Forcing and input observations

Study catchment and data

The Grote Nete is used as study catchment. The available information about the forcing variables (rain and evapotranspiration) and the observed flow where introduced in section 6.2. However, a deterministic value of the observed flow was used to evaluate the model performance in Part III. To include the uncertainty of the observed flows, the water level (stage) measurements are used as well to derive an envelope of expected flow values instead of a deterministic estimate. The derivation based on the rating curve is explained in the next section.

Rating curve uncertainty derivation

The stage-discharge evaluation points of the Geel-Zammel discharge station, represented by triangles in Figure 10.2, are used for deriving the uncertainty on the observations. A power law is assumed to define the relationship between the discharge and the water level:

$$Q = a(h + b)^c \tag{10.1}$$

with, Q the discharge, h the water level and a, b, c fitting parameters.

To estimate the uncertain power law, an uncertainty envelope based on an initial uncertainty estimate of both the discharge derivation and the water level measurements was first defined. By varying the 3 parameters of the power law, those realisations included in the uncertain envelope of the different measurements were used to derive an overall rating curve uncertainty envelope, similar to Pappenberger et al. (2006). A membership function of 1 was given to each rating curve which is within the assumed uncertain boundaries of the measured discharge and water level and zero when outside these boundaries. This method is in line with other methods to assess the rating curve uncertainty that also use sampling based approaches (McMillan et al., 2010) or methods based on fuzzy regression (Westerberg et al., 2011b; Shrestha and Simonovic, 2010).

The same measurement error for all the calibration measurements of the discharge was assumed. Literature reports values between 1.8 % and 8.5 % for discharge

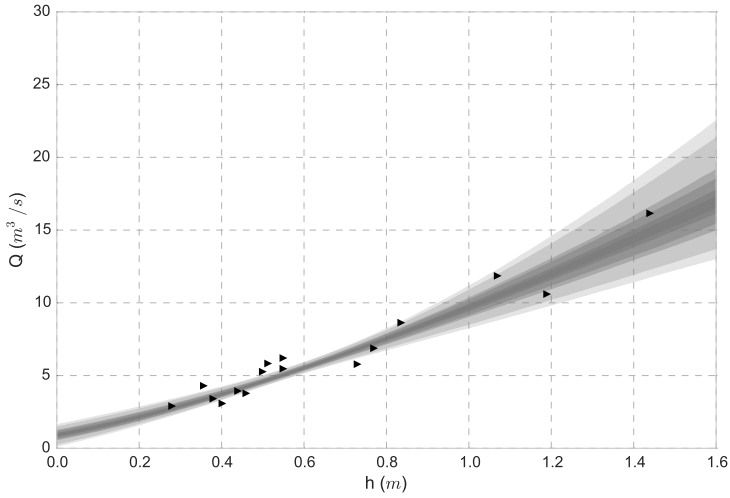


Figure 10.2: Uncertainty outcome based on a 5 % measurement error in Q, the triangles are the measurements and the different gray shades represent different percentiles of the behavioural set of power law realisations

measurements and 3 till 14 mm for the water level measurements (Pappenberger et al., 2006). In this test-case study, a discharge error of 5 % and no error for the water level is assumed. The latter resides in the fact that no specific information about the observation spot was available and that the relative error in the discharge is expected to be significantly larger than the relative error in the water level measurement. More elaborated research would be needed to identify a more reliable value of this uncertainty, since uncertainty in the individual rating curve measurements can be significant for both low and high discharges (Blazkova and Beven, 2009).

When 1 out of 16 membership functions is allowed to be zero (i.e. curve does not cross the defined uncertainty of the observation), the set of behavioural parameter sets can be used to derive uncertainty bounds of the discharge measurements. In this way, the possibility of a very bad measurement is taken into account in the evaluation, without explicitly excluding specific measurements. The resulting uncertainty envelope is shown in Figure 10.2. The uncertainty increases towards lower and higher (extrapolated) values of the stage-discharge measurement points. Since only membership functions of one and zero are used, every behavioural realisation gets the same weight. This assumption is made since the model error was expected to be larger than the measurement error similar to Krueger et al. (2010).

For the same reason, it is not expected that the hydrological model realisations would fall into these measurement uncertainty bounds for all time steps. Thus, adding more detailed information about the observation error structure within the bounds would not add significant information to the model structure evaluation and focus is on the relative differences with increasing uncertainty towards the more extreme values.

For every time step in the flow time series, the measured value was assumed to correspond to the median value in the uncertainty envelope of the rating curve (Figure 10.2). The percentiles of this envelope corresponding to this median value were used to translate the uncertain rating curve into the flow time series uncertainty. The resulting uncertainty bounds from 2003 till 2005 are given in Figure 10.3. These percentiles are used as limits of acceptability in the remainder of the approach (Figure 10.1).

10.5.2 PDM and NAM lumped hydrological model structures

The PDM and NAM model implementations were introduced in chapter 9, both in the original descriptions as in the more generic ODE description. Notice that for this application the original version was used and that the parameter names and symbols in this chapter are according to the original model descriptions of respectively section 9.6.4 and section 9.6.3.

The focus on these two specific model structures is mainly triggered by the relevance for current operational water management in Flanders, since they are models applied in operational water modelling frameworks. The NAM model has been employed successfully to describe the hydrological behaviour of Flemish rivers in the past (Vansteenkiste et al., 2011) and is used by the Flanders Hydraulics Research in their water management activities. Also abroad the concepts and performances of the model structure had been proven adequate in different applications (Refsgaard and Knudsen, 1996).

To screen the parameter space, a brute force sampling approach is used and a total of 500 000 model simulations of both model structures were performed. Sampling of the parameter combinations was performed with a quasi random sampling technique (Sobol, 1967) assuming a uniform distribution between the defined ranges (cfr. section 3.5). Parameter ranges are given in Tables 10.1 and 10.2 for respectively PDM and NAM models. For the PDM model, the parameter ranges are based on those proposed by Cabus (2008). The results of the study performed by

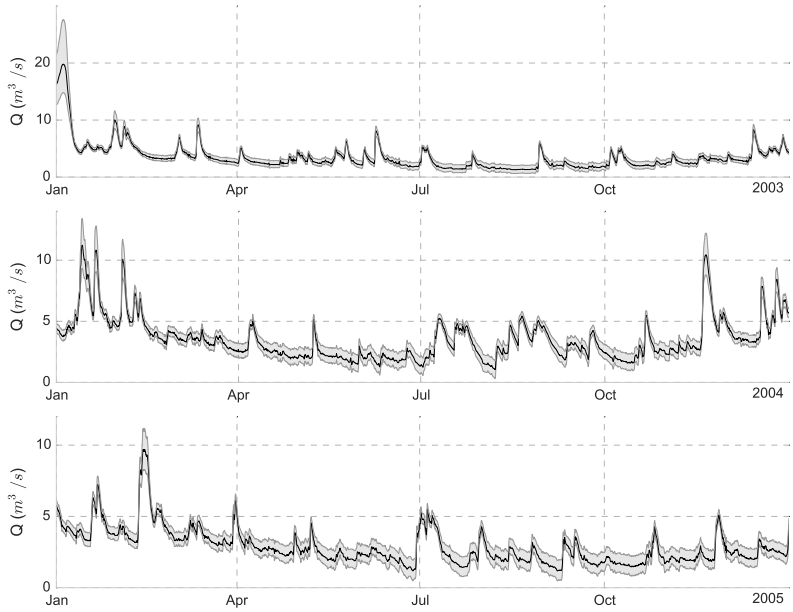


Figure 10.3: Observation uncertainty for the years 2003, 2004 and 2005 with the measured discharges presented as black line. The grey uncertainty band is delimited by the 5th and 95th percentile values as computed from the rating curve analysis. Increasing uncertainties for both lower and higher discharges are apparent. These uncertainty bound are used as limits of acceptibility to assess the model performance.

Vansteenkiste et al. (2011) was used to set up the parameter ranges for the NAM model.

10.5.3 Performance metric: Limits of acceptability

Considering the uncertainty of the measured discharges, the limits of acceptability approach provides a method to define a performance metric that takes the uncertainty into account (cfr. section 3.4.3). The limits of acceptability are directly derived from the uncertainty bounds coming from the rating curve uncertainty. As such, the specified minimum ($Q_{\min}$) and maximum ($Q_{\max}$) limits of acceptability (Figure 10.3) correspond to the data uncertainty. By using these limits, the problem of making assumptions about the statistical characteristics of the modelling

Table 10.1: Overview of the PDM model parameters ranges assumed for the Nete case, based on the ranges proposed by Cabus (2008).

Parameter		Description	Sampling range
$C_{\max}$	(mm)	Maximum storage capacity	160–5000
b		Exponent of Pareto distribution	0.1–2.0
b_e		Exponent in actual evaporation function	1–4
b_g		Exponent of recharge function[a]	/
k_g	($h\,mm^{b_g-1}$)	Groundwater recharge time constant	700–25 000
k_b	($h\,mm^2$)	Base flow time constant	0.0002–1.0
k_f	(h)	Time constants of cascade of two linear reservoirs	0.1–40
S_τ	(h)	Soil tension storage capacity	0–150

[a] value of b_g was set to 1 for all simulations, see section 9.6.4

error needed in Bayesian applications is avoided (Beven et al., 2008; Beven and Freer, 2001; Vrugt and Robinson, 2007).

In order to rank the different model simulations, these limits were translated into a model evaluation score. A similar approach as Westerberg et al. (2011b) and Liu et al. (2009a) is chosen, i.e. the score is -1 and 1 when simulated discharges are equal to respectively the lower and upper limit of the uncertainty bounds and linearly interpolated values are used in between the boundaries (Figure 10.4, left). Summing the absolute values of the scores of the individual time steps results in an aggregated score for each model simulation.

In principle, a model prediction will be selected if all modelled values fall between the specified minimum ($Q_{\min}$) and maximum ($Q_{\max}$) limits of acceptability for all time steps. However, under these criteria, all model realisations are rejected and, similarly to (Blazkova and Beven, 2009; Liu et al., 2009a), relaxation of the criteria is to be considered . A first option is relaxing on the number of observation points that need to satisfy the specified limits. This needs careful verification in order to avoid that periods of non-compliance with the limits, are the periods of interest. A second option is relaxing on the initially set acceptance limits of the individual

Table 10.2: Overview of the NAM model parameters ranges assumed for the Nete case, based on the study performed by Vansteenkiste et al. (2011).

Parameter		Description	Sampling range
$U_{\max}$	(mm)	Maximum water content in the surface storage	3–25
$L_{\max}$	(mm)	Maximum water content in the lower zone	50–250
CQ_{OF}		Overland flow runoff coefficient	0.01–0.99
T_{OF}		Threshold value for overland flow recharge	0–0.7
T_{IF}		Threshold value for inter flow recharge	0–0.7
T_{G}		Threshold value for groundwater recharge	0–0.7
CK_{IF}	(h)	Time constant for inter flow from the surface storage	100–1000
$CK_{1,2}$	(h)	Time constant for overland flow and inter flow routing	3–48
CK_{BF}	(h)	Time constant for base flow routing	500–5000

observation points and thus accepting time steps with scores larger than 1 or smaller than -1 (Liu et al., 2009a). The type of relaxation used in each step in the methodology is explained in the next two sections. Finally, the compliant subset of simulations could also be determined by taking a percentage of best performing simulations considering their summed score.

The scores themselves are both used in the DYNIA and the GLUE approach in order to:

- Make a first subset selection of simulations to apply the DYNIA approach, using the entire calibration period (see section 10.5.4).
- Evaluate and select the simulations in the different time frames set by the DYNIA approach (see section 10.5.4).

- Select the behavioural simulations to derive the prediction uncertainty according to the GLUE approach. For the latter, the scores need to be transformed into weights. The weights of the different behavioural simulations are subsequently used in the GLUE methodology to derive the prediction limits of the ensemble of model realisations (see section 10.5.5 and Figure 10.4).

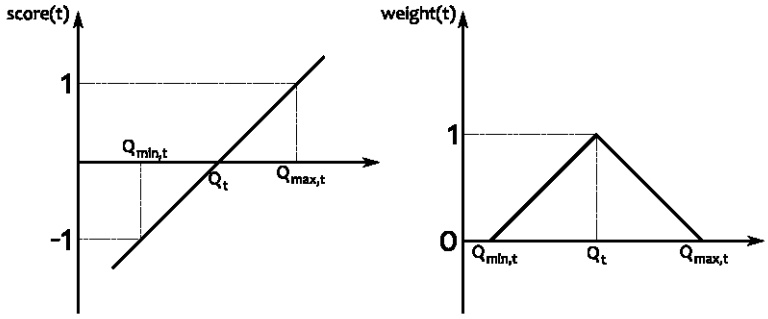


Figure 10.4: Calculation of the scores (left) and weights (right) based on the uncertainty ranges derived from the measured flow. $Q_{\mathrm{min,t}}$ and $Q_{\mathrm{max,t}}$ are the lower and upper limit for the flow uncertainty at time step t and Q_t the measured flow, corresponding to the median of the uncertain measurements. A score of 0 is assigned to simulated values equal to Q_t, -1 to values at the lower limit and 1 to values at the upper limit. Other values are linearly inter- and extrapolated. Scores are converted to weights by a triangular weighting function at every time step. Simulated time steps closer to Q_t receive proportionally higher weights and scores outside the boundaries are 0 in order to construct a likelihood value.

10.5.4 DYNIA approach

Prior application of limits of acceptability

In contrast to Wagener et al. (2003), for this application a two-step application of the DYNIA approach is applied. First, a subset of simulations is selected based on a performance metric aggregated over the entire calibration period. In this way, only those simulations able to satisfy an initial set of limits of acceptability are selected. By only retaining this subset of simulations, the further analysis focuses on the posterior parameter distributions that represent the (overall) dynamics of the system with a certain user-defined minimum level of performance.

After a first rejection of all simulations, a relaxation was applied towards both the limits of the score and the fraction of time steps the simulation needs to be in the allowed envelope. Since over- and underprediction of the simulations is observed at similar degree, both the upper and lower score limits are extended to -2 and 2. In other words, the initially derived uncertainty measurement boundaries seemed to be too conservative. Next to this, the percentage of time that the simulations are allowed to be outside the score limits is set to 10 % of the simulation period. As such, the limits of acceptability are relaxed and a total of 477 parameter combinations for the NAM model and 389 parameter combinations for the PDM model are accepted. The relaxation is done specifically for this case based on expert judgement and should be reconsidered when more or less confidence in either the model structure or the data exists.

Given the applied relaxations, it is important to understand at what time instants the model simulations are violating the score boundaries in order to observe potential systematic failure of the selected simulations. This check was done visually based on an empirical cumulative distribution of the scores over the different time steps. A balance in the number of over- and underpredictions is required for further analysis. Besides the calibration period as a whole, a more detailed check was done on selected periods of the hydrograph. First, a separation was performed to discriminate different modes of the hydrograph similar to Boyle et al. (2000); Wagener et al. (2001a); Krueger et al. (2010). A segmentation was done between driven (wetting up, positive slope of the hydrograph) and non-driven (draining, negative slope of the hydrograph) periods, illustrated in Figure 10.5. A further separation of the non-driven periods in quick and slow non-driven periods was made using a threshold for flow. This threshold was set to the mean flow of the season the period belongs to (in contrast to Wagener et al. (2001a), using overall mean flow) in order to better adapt to the seasonal variations. Secondly, a separate seasonal segmentation was done to evaluate the seasonal effects.

DYNIA application

The DYNIA approach, initially developed by Wagener et al. (2003), is essentially a dynamic extension of the RSA (Hornberger and Spear, 1981) (section 5.8).

An important decision in the analysis, is the selected time window for which the scores are aggregated for each of the parameters, enabling the identification of important response modes. Since for all parameters, the same (uncertain) flow time series is used, a classification between a short (1-5 d), median (5–30 d) and long (30–45 d) window was used for parameters that are expected to mainly contribute

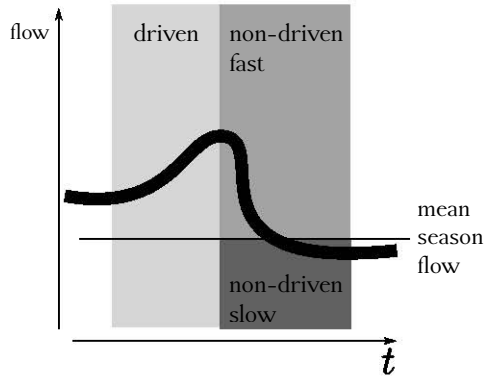


Figure 10.5: Illustration of the segmentation to seperate driven and non-driven periods of the hydrograph, inspired by Boyle et al. (2000). Driven periods are identified by increasing flow values due to incoming rain, whereas non-driven periods are characterised by decreasing flow values. A further distinction is made between non-driven fast periods and non-driven slow periods using on a threshold value, i.e. is the mean flow of the season.

to respectively overland flow, unsaturated zone and groundwater processes. When the window size is too narrow, the influence of the data error could become too influential, whereas too wide window sizes can result in aggregation of different periods of information (Wagener et al., 2003).

By adapting the time frame manually within the proposed ranges, the period that gave the most (visual) information about a parameter's behaviour was selected. Depending on the window size, the time steps at the beginning and end of the time series that are distorted need to be excluded for the interpretation (Wagener et al., 2003). For each parameter of both model structures, a plot was made representing the dynamic identifiability of the parameter.

10.5.5 Prediction uncertainty derivation with GLUE

GLUE (Beven and Binley, 1992; Beven and Freer, 2001) is explained in section 5.9 in chapter 4. It accepts all simulations satisfying the defined requirements and combines them into output variability (uncertainty) limits based on their corresponding performance metric values (Beven, 2006). Hence, it provides insight in the variability of the output under the specific selected conditions of the parameter conditioning process. In the remainder, the output variability will be referred to as prediction uncertainty, considering the common terminology in literature. Notice the discussion in section 5.9 about the legitimacy of referring to uncertainty.

In this application, all model realisations having a model output within the minimum and maximum limits for a sufficient amount of time steps, considering the applied relaxations, are considered as behavioural. The definition of prediction percentiles requires a likelihood weight to be specified for every model run (Beven, 2006). To obtain this aggregated likelihood weight, the scores at all individual time steps are first translated using a triangular weighting function, similar to fuzzy membership functions (Liu et al., 2009a; Westerberg et al., 2011b; Blazkova and Beven, 2009) and then summed up to derive a single weight associated with the particular model realisation, similar to Liu et al. (2009a). Again, other methods to combine the weights of the individual points are possible (e.g. giving periods of low flow and high flow more importance) and worth testing, which fits in the performance metric approach. Models that produce flow predictions close to the observations will have higher weights and vice versa. Other conceptualisations about the measurement error could be used to construct these weights as well.

To derive the prediction uncertainty, the same limits of acceptability as those of the first subset selection (section 10.5.4) were used in a first analysis. As such, the information about the parameters and structures can be related to the prediction uncertainty coming from this set of selected simulations. Subsequently, a second (additional, i.e. result of the DYNIA application, see section 10.6.1) selection of behavioural parameter sets was done based on the scores during the individual seasons. In this way, the analysis is based on a seasonal segmentation in contrast to the analysis of the entire calibration period. Limits of acceptability were put for each season separately with score limits of -2.5 and 2.5 and a maximum percentage of 5 % of the time steps that these limits might be trespassed. As such, less concern is put on the individual scores, but more on the percentage of time in contrast to the selection criteria for the entire period (aggregated scores). This sub period posterior parameter evaluation was performed to get insight in the behaviour of the model structures with respect to the seasonal dynamics.

10.6 Results

10.6.1 DYNIA model evaluation

Prior application of limits of acceptability

Figure 10.6 shows the scores for the entire calibration, as well as for the driven and non-driven periods. 90 % of the time steps are within the -2 to 2 boundaries defined as (relaxed) limits of acceptability. The gray bounds indicate the -1 and 1 boundaries: they indicate those model simulations with a prediction within the derived observation uncertainty bounds. No unbalanced over- or underestimation of the scores is observed, except for a slight skewness of the scores during the non-driven slow periods. This indicates shortcomings of both the model structures in representing the long-term drying up of the catchment. Based on the seasonal scores (Figure 10.7), differences between both models are clearer and the long term seasonal limitations appear to predominate the short term representation of the wetting and drying after a rain event. Furthermore, larger differences in the histogram plots in between the models indicate the mutual difference between the model structures to be more apparent at the seasonal level.

The imbalance between over- and underestimation of the scores is not excessive and the relaxation of the score is wide enough to minimize the risk of type II errors, i.e. excluding potential accurate simulations. As such, this subset selection is considered sufficient to initiate further DYNIA analysis. The approach was used to focus on the selected parameter sets and to augment the insight in the uncertainty inherent to model structures.

Results of DYNIA for NAM model

Figure 10.8 shows the identifiability analysis for the parameter (T_{OF}), which is the threshold value for overland flow of the NAM model. The plot visualizes both the DYNIA results in the parameter-time space and the derived IC over time. The range of the y-axis at the parameter side is taken from the original parameter boundaries. The combined analysis allows on the one hand to identify periods with high identifiability and on the other hand to verify the location of optima in the parameter space during these periods. The IC of T_{OF} is the highest during summer rain events, where the width of the confidence limits is decreasing and the confidence region is centered around lower values of the parameter. The

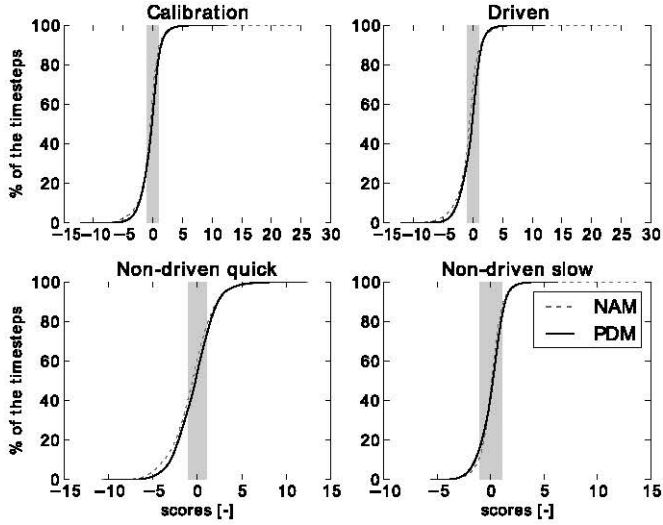


Figure 10.6: Empirical cumulative distribution of the scores of all behavioural model realisations for the entire calibration period as well as selected parts of the hydrograph for both NAM and PDM model. The histograms are normalised by the number of behavioural simulations and represent the % of time steps of the defined period. The gray band represents the -1 and 1 boundaries of the measured uncertainty.

mode of the distributions of the parameter value fluctuates during the remaining periods without particular optima, indicating that varying values of the parameter can yield similar score values. This can be explained either by the influence of the other parameters in the model (i.e. identifiability problems) or by the model output being not sensitive to this parameter during these periods. However, during the summer months the threshold is more identifiable and has generally a lower value compared to the other periods (generating more overland flow).

The $L_{\max}$ parameter representing the maximum water storage in the lower soil between root zone and groundwater (Figure 10.9) evolves towards different parameter values during different periods. Lower values appear during winter months in 2004 and 2005, whereas higher values are obtained during spring months of 2003 and 2004. As stated by Wagener et al. (2003), this typically indicates a failure of the model structure due to the inconsistency in the way the model fits the observed flow during different seasons. Moreover, parameter CK_{BF} behaves in the opposite direction to compensate for this seasonal variation (results not

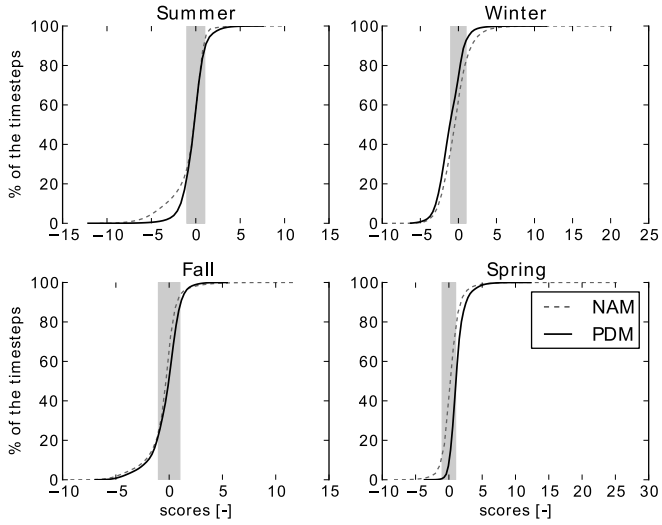


Figure 10.7: Empirical cumulative distribution of the scores of all behavioural model simulations for the different seasons for both NAM and PDM model. The histograms are normalised by the number of behavioural simulations and represent the % of time steps of the defined period. The gray band represents the -1 and 1 boundaries of the measured uncertainty.

shown, but the seasonal parameter switch is also visible for the winter season in Figure 10.14).

Similar analyses of the other parameters shown in appendix A of the NAM model show a shift towards very low values of $U_{\max}$ during certain rain events, but this causes at the same time overestimation of the peaks. CQ_{OF} is identifiable during winter events and also for T_{G} seasonal variation of the optimal parameter value is recognisable, but not as distinct as for the previous parameters. For T_{IF}, identifiability of the parameter is low throughout the entire calibration period, whereas for CK_{IF} a small shift towards higher values is observed in winter months when the catchment is in wet condition. Differences in the area of identifiability of the $CK_{1,2}$ parameter during rising and falling limbs indicates that using the same time constant for overland flow and inter flow will be too simplistic to capture the retention of the basin.

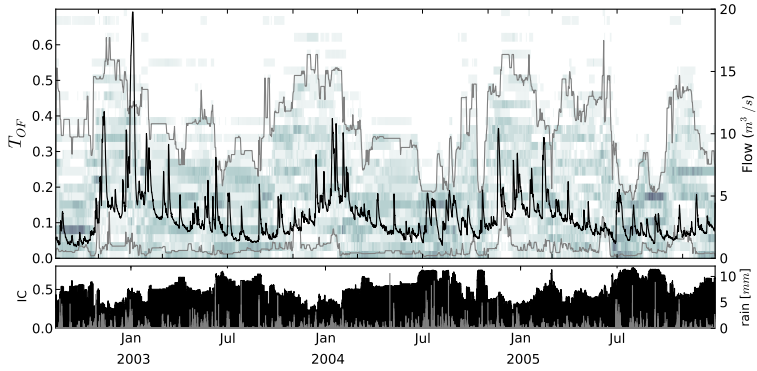


Figure 10.8: Results of the DYNIA procedure for parameter T_{OF} (NAM model) applied to the behavioural model simulations for the entire calibration period. The black line in the top graph shows the measured streamflow (right axis). The dark gray lines are the 90 % confidence limits derived from the cumulative distribution of the parameter values (left axis) of the behavioural model realisations and the gray shading indicates the size of the gradient of these distributions, with a darker color for a higher value (better identifiable). A time window of 3 days was used since the parameter belongs to the group with quick response processes. In the lower graph the rain is shown in gray (right axis) together with the Information Content (IC; black , left axis), defined by one minus the width of the confidence limits over the parameter range. Identification of T_{OF} is mainly possible during summer storms.

Results of DYNIA for PDM model

For the PDM model, Figure 10.10 shows the dynamic analysis of the maximum storage capacity ($C_{\max}$). In this case, the periods with the highest information content along the entire period are the periods of heavy rain. In these periods convergence towards clearly defined parameter ranges is much more present than in other periods. In the recession after the winters of 2003 and 2004, a shifting towards higher values together with a decrease in identifiability is visible, but to a lesser extent than the shifting of the $L_{\max}$ parameter of the NAM model (Figure 10.9), indicating a better representation of the seasonal variation in the catchment.

The parameter b that defines the shape of the pareto distribution and thus represents the spatial variation in the catchment, is the second parameter defining the unsaturated zone processes (Figure 10.11). During most of the year, parameter b

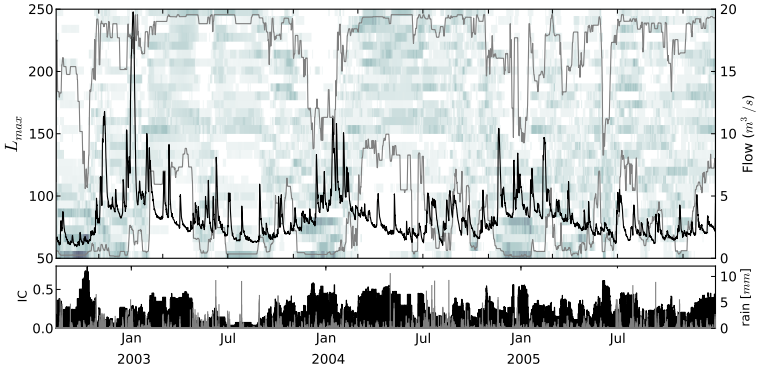


Figure 10.9: Results of the DYNIA procedure for parameter L_{max} (NAM model) applied to the behavioural model simulations for the entire calibration period (see Figure 10.8 for explanation). Changing regions of identifiability are identified in summer and winter, possibly indicating model structural shortcomings.

strives to lower values, except for the spring periods, where the parameter is less identifiable, probably due to the interaction with $C_{\max}$. Furthermore, the increase of the parameter value indicates more variation in the catchment in terms of soil storage availability.

Similar plots of other parameters of the PDM model are shown in appendix A. The exponent of the evaporation function b_{e} does not show a distinct area of identifiability. The groundwater recharge constant k_{g} is much more identifiable than the base flow time constant k_{b}, showing the importance of the drainage function to capture the seasonal variation of the groundwater. The storage capacity S_{τ} of the drainage function on the other hand is less identifiable, whereas the routing of the overland flow (k_{f}) is identifiable during the entire period, but exhibits jumps between two optima that are not directly seasonally related.

10.6.2 Prediction uncertainty derivation with GLUE

Prediction uncertainty

Based on the set of accepted parameter combinations and their corresponding (normalised) weights, the output prediction uncertainty of both model structures is computed. Figure 10.12 gives the uncertainty (90 % prediction uncertainty) for 2004 and compares the uncertainty about the observations with the modelled

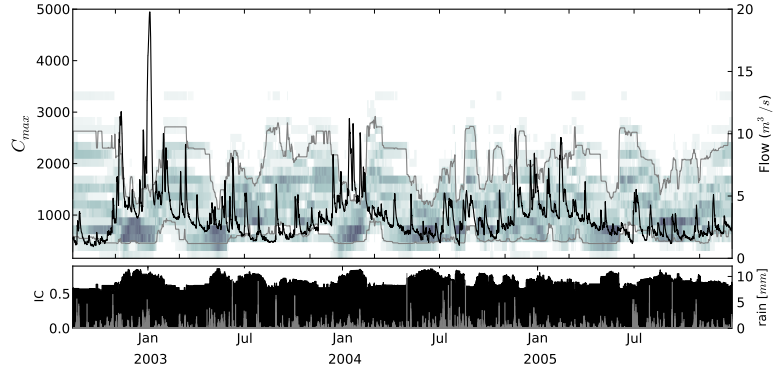


Figure 10.10: Results of the DYNIA procedure for parameter C_{max} (PDM model) applied to the behavioural model simulations for the entire calibration period (see Figure 10.8 for explanation). Identifiability is the highest in periods of heavy rains with a consistent tendency towards values of about 700 mm.

prediction uncertainty for 2004. PDM tends to underpredict the peaks during winter months, but captures the dynamic behaviour in the summer months. The variation in June is completely missed by the NAM model. Both models are overestimating the flow peaks in March. Mainly the periods where one out of the two models is unable to predict the dynamics are useful to distinguish model structural differences.

For the validation period, the prediction uncertainty of 2006 is shown in Figure 10.13. Similar differences between the model structures as compared to the calibration period can be observed. PDM better captures the recession periods in July and October and the NAM model predicts in general higher peak discharges. During storms, uncertainty boundaries related to the NAM model are wider compared to those of the PDM model. The similarity in the failures of the models in both calibration and validation periods further confirms that the conclusions of the identifiability analysis are independent of the choice of calibration period.

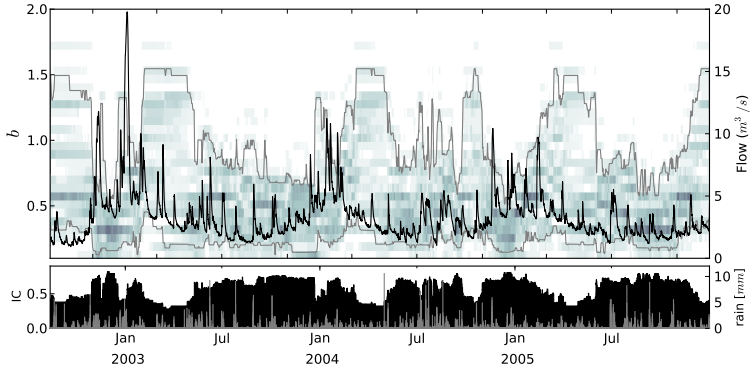


Figure 10.11: Results of the DYNIA procedure for parameter b (PDM model) applied to the behavioural model simulations for the entire calibration period (see Figure 10.8 for explanation). Higher optimal values during winter and spring months increases the overland flows.

Posterior evaluation of periodically selected parameter combinations

In Figure 10.14 the posterior parameter distributions of the NAM model are shown for each season. The resulting posterior parameter distributions are in correspondence with the model identification (section 10.5.4). Seasonal variation of optimal parameter values is mainly visible for parameters L_{max}, CK_{BF} and CK_{IF}. Overland flow parameters, CQ_{OF} and $CK_{1,2}$, are highly identifiable during winter, whereas T_{OF} is during summer months. Nevertheless, seasonal differences are visible due to rain events happening during respectively wet or dry conditions of the catchment. The posterior distributions of the parameter T_{IF} do not contain a small, clearly defined optimal region in any season. Since also the DYNIA approach revealed no specific region of identifiability, the usefulness of the infiltration threshold for this application can be questioned and simplifying the inter flow description (leaving out the T_{IF} parameter) can be considered.

The posterior distribution of the parameters of the PDM are shown in Figure 10.15. The seasonal variation that is visible for the $C_{\max}$ parameter (mainly winter and fall) in combination with parameter b is different to the DYNIA results since the higher posterior values during winter were not accepted in the limits of acceptability set for the entire calibration period. A higher value for $C_{\max}$ (more water storage capacity) would help the prediction during winter but tends to predict the rest of the hydrograph wrongly. The main differences with the seasonal variation is

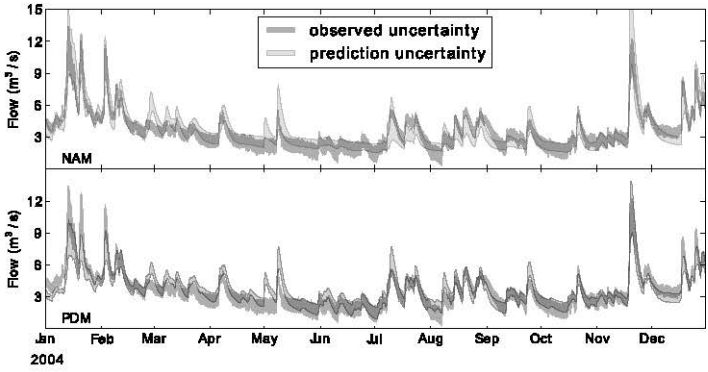


Figure 10.12: Uncertainty boundaries for measured and predicted flow during 2004 (calibration), both confined by the 5 % and 95 % percentiles. PDM tends to underpredict the peaks during winter months, but captures the dynamic behaviour in the summer months. The variation in June is completely missed by the NAM model.

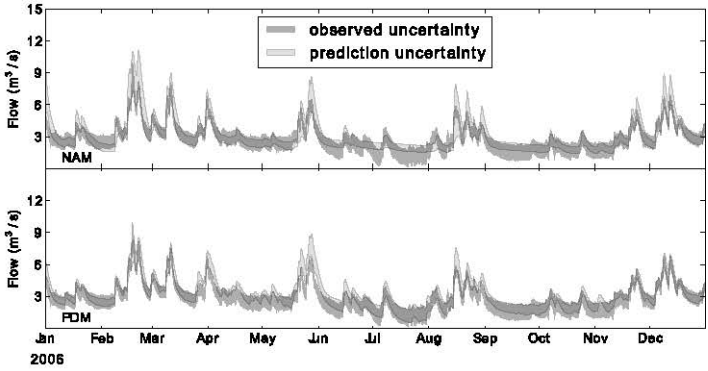


Figure 10.13: Uncertainty boundaries for measured and predicted flow during 2006 (validation), both presented by the 5 % and 95 % percentiles. glspdm better captures the recession periods in July and October and the NAM model predicts in general higher peak discharges. During storms, uncertainty boundaries related to the NAM model are wider compared to those of the PDM model.

noticeable for parameter k_g. Again, these high values in summer and spring were not taken into account in the DYNIA approach. These high values decrease the drainage towards the groundwater reservoir. The posterior of the non-driven slow

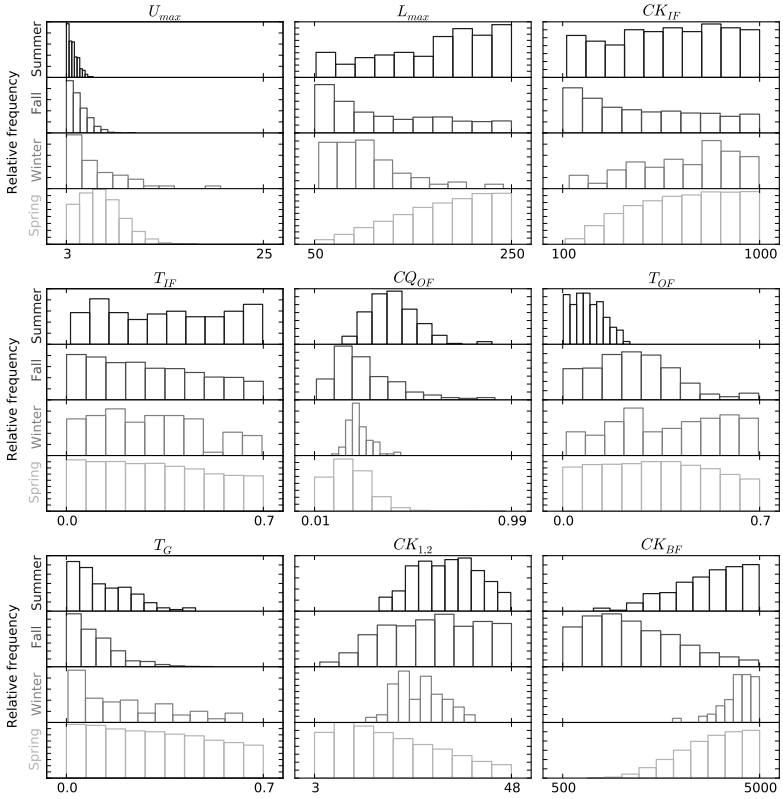


Figure 10.14: Posterior parameter sets of the behavioural model simulations selected based on the specific part of the hydrograph for the NAM model. Driven periods and non-driven quick periods are excluded since no behavioural sets were present according to the used limits of acceptability.

supports the convergence towards winter values. Based on the seasonal selection k_b and S_τ are not identifiable.

10.7 Discussion

This chapter combines for the first time the limits of acceptability approach (Beven, 2008b) with the dynamical identifiability approach DYNIA (Wagener

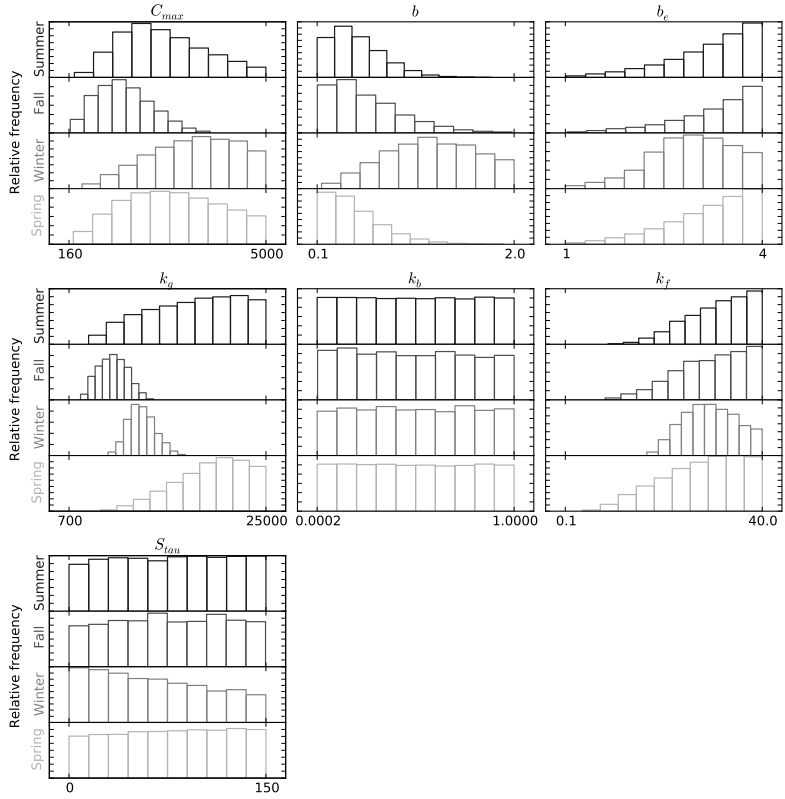


Figure 10.15: Posterior parameter sets of the behavioural model simulations selected based on the specific part of the hydrograph for the PDM model. Driven periods and non-driven quick periods are excluded since no behavioural sets were present according to the used limits of acceptability.

et al., 2003). By doing this it is possible to evaluate the potential of detecting model structural deficiencies, while taking into account the rating curve uncertainty. Using the resulting uncertainty band of the flow time series as evaluation limits, one does not need to make explicit assumptions about the nature of the modelling errors (Beven, 2008b). When the analysis of the obtained evaluation scores for different subperiods is lacking clear indication of over- and underprediction (Figure 10.7), the added value of the DYNIA approach becomes clear. Indeed, by applying the DYNIA approach, one can get insight into the model structural

limitations. Comparable information about the parameter time-variation is derived by the subperiod parameter selection (section 10.6.2), but this is based on the knowledge of seasonal defects brought by the DYNIA approach. This, in combination with the ease of use, illustrates the advantages of applying the DYNIA approach as generic information source for model structure evaluation and improvement in comparison to model evaluation based on a comparison of the performance towards one or more performance metrics.

A first difference between the applied models is the soil moisture storage component. NAM is using one upper and lower storage reservoir, whereas PDM uses the probability distribution concept aiming at conceptually introducing the spatial variability. Furthermore, a linear routing of the groundwater is used in the NAM model in contrast to a non-linear routing of PDM. Groundwater recharge is comparable when b_g is assumed 1 for the PDM model. The differentiation in 3 subflows in the NAM model, against 2 in the PDM model is partly compensated for by the use of one time constant for both overland flow and inter flow in the NAM model. The limitations to simulate the seasonal dynamics are dominating the peak discharges of the individual rain events, mainly dominated by the soil moisture storage conceptualisation. From the results presented here, the probability distribution approach from the PDM model seems to be more suited.

Moreover, capturing the seasonal dynamics is in this catchment mainly related to the groundwater representation, due to the sandy soils and low slopes in the catchment. The absence of identifiability of the PDM base flow time constant (k_b) and the interplay of the seasonal variation in the NAM base flow time constant (CK_BF) with the soil moisture L_max suggest shortcomings for both models, albeit for different reasons.

In the NAM model, a clear compensation of the parameter values suggests that the inability of the soil moisture storage is causing these problems, probably due to the inability to capture the dynamics by only one reservoir. During winter months, lower L_max values produce more runoff in combination with higher base flow time constants to prevent the overprediction. After the winter months, higher L_max values are needed to decrease the flows together with lower base flow routing. In general, the combination of the small U_max reservoir and the single L_max reservoir accounting for unsaturated zone is not sufficient to incorporate seasonal variations. The insufficiency of the unsaturated zone concept of the NAM model to capture the water retention in the catchment throughout the year is further supported by the seasonal variation of the posterior parameter distributions as pointed out in section 10.6.2.

In the PDM model, the seasonal variation is mainly captured by the soil moisture variation in combination with the identifiable recharge parameter (k_g), inducing the limited influence of the k_b parameter. Higher values of b indicate a higher spatial variation during winter months (indicating the shortcoming of a single reservoir based model structure), whereas the low values during the rest of the year suggest more uniformity in the catchment implying that a single storage may be sufficient during these periods. McMillan et al. (2011) reached similar conclusions based on the non-uniqueness of the storage-discharge relationship, suggesting that multiple reservoirs are required. As such, the seasonal variation is captured by varying proportions of flow from the different reservoirs (cfr. the PDM approach).

Since the DYNIA approach starts from the simulations selected by the same limits of acceptability as the GLUE approach, the characteristics of the predictions can be compared. The mismatch between the flow predictions of the NAM model in the falling limbs was observed by applying the GLUE approach and can be explained by the changes in the region of identifiability of the $CK_{1,2}$ parameter, which is shown by the DYNIA approach. The overestimation of the peaks and their larger prediction uncertainty in the NAM model is mostly related to the lack of identifiability of the treshold T_{OF} and thus related to the influence of the lower zone configuration (see Equation 9.7). Notice that the GLUE method is actually used as a sensitivity analysis on the output variability referred to in section 5.9.2, comparing the effect of the defined limits on the output variability of two models. The output uncertainties should be regarded in this way and only interpreted relative to each other and to the observation uncertainty.

Foregoing conclusions are made based on the application on one single catchment and might be different for other catchments with different specifications. The method can be applied to any model structure analysis and type of hydrological data. To derive generalised conclusions, a larger effort using a larger set of basins would need to be used. However, this was beyond the scope of this dissertation that merely wanted to demonstrate the methodology and its assets.

A restriction in the application of the limits of acceptability approach is the need for relaxation of the initial limits of acceptability to avoid rejection of all model simulations (both in terms of parameterisation and structure). Similar relaxations were also needed in the work of Blazkova and Beven (2009) and Liu et al. (2009a).

However, since the focus is on model evaluation, the approach is rather based on rejection of bad parameter sets and model structures than on parameter optimization (moreover, the number of simulations would be far too insufficient to identify the overall optimal region). For model evaluation, it is important to identify and

focus on particular parts of the time series that are not well simulated (Beven, 2008b). This *learning by rejecting* is made possible by consecutive relaxing of the limits of acceptability. In the presented approach, 2 major degrees of freedom were altered. The first one is the % of allowed time trespassing the limits and the second one is relaxing of the limits.

When putting rigorous requirements on the % of time and at the same time relaxing the limits, more focus is given towards the prediction of the general behaviour of the dynamics. Alternatively, more focus can be put to periods of violating the initial derived limits by relaxing the % of time, and keeping the original limits. In the application here, a relaxation of both was used to gain general insight in the behaviour of the resulting behavioural simulations. This choice will, however, be case specific. It depends on the expected uncertainty in the data and the confidence in the model structures to be tested.

The described relaxations were taken into account in both the DYNIA and GLUE approach. The resulting behavioural model simulations used in section 10.5.3 were selected based on a time-aggregating performance metric, whereas in section 10.6.2 separate limits on different response modes of the hydrograph are used. However, comparable results were obtained by Peters et al. (2003). The DYNIA approach allows evaluating the selected simulations in function of time. In model evaluation, the use of multiple, non-commensurable, evaluation functions focusing on different underlying assumptions is essential (Gupta et al., 1998; Winsemius et al., 2009), but the selection of the most appropriate criterion is not always straightforward. The application of the DYNIA approach can give guidance in the selection of performance metrics. For this application example, the use of a total seasonal volume could support the model optimization for practical applications. Besides, by focusing on the behavioural simulations with DYNIA, information is extracted about the reasons for the lack of identifiability of the selected (behavioural) parameter sets. Insight is given in how identification (in terms of parameter space and model structures) can be improved, leading to a more objective and guided reasoning when defining limits of acceptability.

In summary, incorporating the DYNIA approach in the model structure analysis methodology (Figure 10.1) is a straightforward way to discover potential pitfalls and to enhance the learning curve about model structure improvement. Looking into model performance in function of time gives guidance towards model optimization and identification. By incorporating the discharge uncertainty, potential periods of wrong measurements (so called *disinformative observations* in Beven and Westerberg (2011)) are less influential on the model evaluation, making it less biased compared to using deterministic flow values. Since these disinformative pe-

riods lead to biased inference of the parameter distributions (Beven et al., 2011), these periods are indicated by the DYNIA approach and can be further checked for. Furthermore, it is shown that the uncertainty about the observations is not inhibiting the identification of deficiencies in model structure. Still, the use of erroneous input forcing (i.e. rainfall and evapotranspiration data) can obscure the differences in model performance. Accounting for input forcing errors in the structure evaluation can potentially clarify parameter value switches (Kavetski et al., 2006b,a; Vrugt et al., 2008a).

From a practical point of view, the modeller has different options facing these structural flaws:

- Model rejection can be the conclusion, given rise to model adaptations or developing new ones. Since the method offers knowledge about where the model fails, a starting point for model structure adaptation is inherently suggested by the method. Bringing in more physical based reasoning can be the conclusion as well as simplifying the current model. More physical reasoning is needed when physical processes are missed, whereas simplification is needed when overparameterization is the case.
- When different structures are acceptable and their imperfections are complementary (meaning they have shortcomings for different reasons), the modeller can bring the results together in an ensemble. When different model structures do have common pitfalls, the incorporation of both is redundant.
- When the results suggest disinformative observation periods instead of model structural failures, the modeller needs to further check the potential errors in the discharge records.
- The time-dependent information assists the modeller in selecting a representative set of objective functions for further model assessment. Selecting objective functions focussing on the 'potential' pitfalls of the model structure is of more use than an overall Nash-sutcliffe or RMSE function.

The workflow applied is believed to be more generic in use than the illustrative case described in this chapter. Data different from flow measurements, such as groundwater level information or isotope data (Fenicia et al., 2008; Winsemius et al., 2009) can be used in addition to derive extra limits of acceptability. However, in many cases these types of data are not available and the flow time series remain the main basis for model evaluation. Finally, also other model structures can be incorporated in the analysis or the information of the identifiability analysis can be used to propose model structure adaptations.

10.8 Conclusions

This chapter combined different methods implemented in the pystran Python Package 4 of chapter 3 to identify and explain model structure deficiencies. The application is done on two specific instances of the pyFUSE model environment of chapter 9, NAM and PDM, because of their applications in current operational water management. As such, it illustrates how the entire set of modelling tools presented in previous chapters can be combined to propose strategies for model evaluation. More specifically, the application of this chapter contributes to (1) an improved model structure evaluation, preventing the modeller from making type I and type II errors and (2) gain insight in the derivation of the prediction uncertainty.

It starts from the idea of using limits of acceptability, both by the rating curve application and by the ability to propose evaluation functions that are able to discriminate the model structures on their performance. The latter information comes from the DYNIA approach, which indicates where model structures have potential pitfalls. Instead of testing multiple objective functions hoping that differences will be seen, the DYNIA analysis instantly indicates *where* these differences can be found. Practically for the presented analysis, the seasonal evaluation is essential to compare the performance of both models. Parameter identification is directly evaluated by the DYNIA approach, which provides a direct generally applicable strategy to identify model structure failures. As such, the usage of temporal parameter identification methods is still a promising technique for model structure evaluation.

Still, the DYNIA method suffers from the subjectivity in the relaxation of the limits of acceptability and the user-defined moving window for which the scores are aggregated. To overcome these limitations, a new method, called Bidirectional Reach (BReach) (Van Eerdenbrugh et al., 2016a,b), is currently developed which adopts the idea of a time step based model evaluation but overcomes the subjective relaxations by combining the information of multiple relaxation levels. Moreover, by checking the distance for which a parameter combination performs according to the limits and relaxation (referred to as reach) for each observation individually, the method is independent of a chosen window size. Hence, the BReach method overcomes the major drawbacks enlisted.

PART V

EPILOGUE

CHAPTER 11

General conclusions

When a proper mathematical model is available, it becomes a powerful tool for both scientists and engineers. It enables to evaluate the process behaviour under a variety of different conditions both rapidly and inexpensively. Moreover, different *what if* scenarios can be tested without the need of influencing or disturbing the actual process itself, which is crucial in an environmental context.

Modelling is well-developed in a wide range of scientific disciplines. More specific, the group of continuous dynamical models, generally described by a set of ODEs, is frequently used in a wide range of existing environmental modelling environments and applications, although sometimes hidden from the end-user within the (monolithic) implementation.

Within any modelling exercise, the system to describe needs to be defined. The system is the part of reality that is being studied and always depends on the research question at hand. In environmental modelling, a wide range of spatial and temporal scales is possible (cfr. bacterial activity of a reactor versus climate models). The model represent a conceptual representation of the system as a set of process descriptions, i.e. mathematical equations.

Environmental science deals with complex structures characterized by many interacting processes and the representation in model equations is always a simplified version of the real system. The identification of a proper mathematical model is a learning process just as any kind of scientific investigation and is prone to falsification. In other words, we learn about the system behaviour by failing to represent it.

Each system is unique. The environment itself is highly heterogeneous and the availability of observations as well as the modelling purpose itself is case- and system dependent. Hence, it is clear that a tailor-made approach is essential to cope with these variations.

Nevertheless, some hampering factors were identified in a first stage of this dissertation. These factors provoke a conservatism in the modelling field. Many environmental modelling studies are limited to the tweaking of model parameters, which is insufficient regarding the requirement for customization.

11.1 Observed conservatism in modelling

In Part I, some main bottlenecks were identified and discussed in more detail. First, the incoherence in the terminology, notwithstanding the similar mathematical framework, hampers collaboration, makes coherence lacking and eliminates the confidence of stakeholders and practitioners.

Secondly, the quest for the *ultimate super* model drives model development towards increased detail of the process descriptions averse to the necessity of sufficient data, required to test these detailed hypotheses. This gives rise to an identifiability problem, where it becomes impossible to distinguish alternative hypotheses (representations), and limits the testability of models.

The latter is enforced by a third factor of protectionism towards the own (model) creation and the related bias towards positive reports focusing only on the capabilities of the proposed model structure.

A fourth identified factor arises from the classic approach of model software design. The direct impact of the architecture and implementation is often ignored and models are delivered as closed-source, monolithic entities as an *all in one* approach. With regard to the evaluation process, this limits the ability to attribute differences in model behaviour to the chosen process descriptions. Besides, it leads to redundant implementations, it limits the capability to adopt new insights and causes a general lack of reproducibility.

Fifth, besides continuous remarks from scientists within the different disciplines about inferior model evaluation practices, model evaluation is still regularly limited to the one-liner *the model fits the data quite well*. Indeed, it is true that any model can be falsified under stringent conditions and one should strive to check if the model is appropriate for its intended purpose. However, the uniqueness of each

model study requires also adaptation in the evaluation, which is not provided by using the same aggregated performance metrics over and over again.

Finally, the intrinsic heterogeneity of the natural environment which is an open and uncontrollable system compared to, for example, an industrial setting makes the modelling process more challenging.

As intended by **Objective D.1**, the identification and clarification of these bottlenecks provide an valuable insight for the environmental modelling community.

11.2 The diagnostic approach

To counteract the observed hampering factors, a diagnostic framework for further model development and application was initiated as put forward by **Objective D.2**:

- Accept the idea of **multiple working hypotheses** and consider model structure building (identification) as a learning process based on failures
- To make this practically and technically possible, **flexibility** in model development in an **open and transparent** manner, is a necessary condition
- Extending the scrutiny of **model structure evaluation** is essential in any stage of the model exercise, going beyond current model calibration practices

The acceptance of multiple working hypotheses is a direct answer to the failing quest towards far too detailed model descriptions that cannot be supported by a sufficient set of observations. Any conceptual representation, i.e. model structure, is merely a hypothesis about the system functioning and can be supported or falsified by the available observations. At the same time, this concept provides intrinsically a defence against the protectionism towards any created model structure and diminishes the exaggerated focus of treating a model structure as an end-product.

The pragmatic response to the acceptance of multiple hypotheses, is the provision of flexibility in the model construction and identification process. It was illustrated in section 2.5.2 that flexibility is provided by a wide range of existing software environments and frameworks, however these are not always supporting a rejection framework (lack of transparency, coarse granularity...). **Objective S.1** aimed to derive a set of requirements for model structure development that support the

multiple working hypotheses approach. These requirements were defined based on current literature sources:

- Supporting **alternative representations** of the considered processes
- Provide **alternative interconnections** between model processes and components, i.e. construction options
- A clear **separation in between the mathematical and the computational model**
- **Accessible and modular code** implementations

To fulfil these requirements, the finest granularity was used in the dissertation, i.e. adaptation on the implementation of the ODEs themselves. Actually, any flexible framework that supports flexibility on this level, while keeping the computational model separated from the mathematical model itself, is able to support such an approach, on the condition that openness to the model source code is provided.

The final element of the diagnostic approach is the need for an improved model evaluation in function of the identification process. This is a generalisation of the current calibration procedure towards a combined and iterative process of parameter and process (model structural) adaptation. Practical identifiability, both in terms of parameters and model components, is the guiding principle during the evaluation. This means that model structures should contain influential parameters which effects on the model output are not cancelling each other out. In other words, process descriptions used, should have an identifiable functionality.

Since the available observations are mostly the limiting factor to identifiability, all efforts to extract the utmost information content from the available data should be made. This task is complementary to the search for additional data sets offering new information.

The elements of the diagnostic approach were used as main driving principles in the execution of the remaining of the dissertation.

11.3 Tools to support a diagnostic approach

In the second part of the dissertation, the existing tools for model evaluation were interpreted based on a diagnostic approach. First of all, the requirements in function of current environmental modelling environments were identified, resulting

in the selection of model independent implementations that rely on a numerical approximation and take into account the entire parameter space.

A wide range of tools with these characteristics are described in literature, typically considered within a categorization towards their focus either on model calibration, sensitivity analysis or uncertainty analysis. However, the fragmentation and lack of coherence is apparent, resulting in redundancy in the implementations. Moreover, from a practical point of view, the implementations do not always support an extensive exploration leading to a*default*-setting application with aggregated performance metrics that do not support to differentiate between alternative model representations.

This was counteracted by a **metric oriented approach** (**Objective E.1**), focusing on the construction of multiple aggregation metrics of time series that can be translated to different performance metrics. The resulting (performance) metrics can be called by algorithms for either optimization, sensitivity analysis or identifiability analysis. Besides, a clear separation between the sampling strategy, the metric construction and the algorithmic evaluation itself, reduces the overlap and reveals the common elements in many of the existing methods in literature.

The combination of both time-variant and aggregated metrics by multiple methods for a respirometric model illustrated the central position metrics have, as aimed for by **Objective A.1**. The identifiability and model calibration of a respirometric model structure with an additional time-lag component was verified. The analysis revealed that practical identifiability of the time-lag extension could be confirmed, given the availability of experimental data for which the ratio between the added substrate and the biomass is high enough.

The particular advantages of sensitivity analysis to assess the identifiability of parameters lead to the decision to make a number of existing algorithms for SA available. As intended by **Objective E.2**, the combination of an extensive description of the individual methodologies in combination with the release of the code to effectively apply these methods tries to overcome current lack of transparency in the application of SA methods.

The pool of available methods provides the opportunity to select a SA method that is fit for purpose, keeping in mind the computational effort. This was translated in a flow-chart that guides the user in the selection process. Still, the opportunity of recycling simulations amongst different methods has been highlighted and would provide the opportunity to combine the information provided by existing methods without the need of additional simulations.

The entire set of implementations and methods of Part II do rely on already existing methods described in literature. However, the metric oriented approach puts the focus where environmental modellers need it most: **the ability to translate a specific modelling objective in a set of (performance) metrics that are able to diagnose the model structure**. These metrics can be used to evaluate the appropriateness of the model structure for the objective at hand and considering the available data. By facilitating the link with existing algorithms in a modular framework while providing sufficient theoretical background about the method, the application of sensitivity analysis in a transparent manner is facilitated.

11.4 Application of diagnostic approach to hydrological modelling

In Part III and Part IV, the focus is on the application of the diagnostic approach on hydrological modelling, more specific on lumped hydrological models. As illustrated in the dissertation, this type of model structures can be converted to ODE based model structures.

An existing model environment, i.e. the VHM, is the starting point of the application in Part III. The flexibility in the implementation of lumped hydrological models is further generalised in Part IV. In this section, conclusions are drawn with respect to the applications of the diagnostic approach proposed.

11.4.1 Evaluation of alternative representations within a flexible framework

The rationale of the VHM is the consideration of model structures as flexible entities. Moreover, it considers the model building process as a combined effort of model structure identification and model calibration. As such, the VHM approach is compatible with a diagnostic approach and was selected as a case study. Based on the approach, a set of model decisions was defined and the suitability assessed.

In accordance with the multiple hypotheses requirements, VHM does provide alternative representations of the processes and alternative construction options. However, the original VHM model implementation (source code) is not available. Therefore, inspecting the separation between the mathematical and computational

model was not feasible. To comply to the requirements of the multiple working hypotheses approach, a new openly accessible version was implemented in python and verified with the outcomes of the original implementation.

Ensemble evaluation

The flexibility provided by the VHM approach resulted into a set of 24 different representations of the hydrological catchment, based on four types of model decisions that were chosen for the case study: leaving out either the inter flow component, the (non)-linearity of the soil storage, the dependency on antecedent rain to represent saturation excess overflow and three routing alternatives. Furthermore, it was decided to focus on two separate modelling objectives: respectively the ability to represent either high flows or low flows. A set of performance metrics was chosen, based on the flow duration curve, together with the well-known NSE metric.

The specific focus towards the low flow and high flow performance metrics that are based on the flow duration curve, lead to different optimal parameter combinations, illustrating the potential of tuning a model in function of a specific purpose. Furthermore, the effect of the chosen performance metric on the resulting parameter set was shown, similar to reports of previous authors (Gupta et al., 1998; Boyle et al., 2000).

The main cause is the lack of identifiability, leading to a wide range in parameter combinations able to achieve a comparable performance. So, when applying these models in an operational setting, the application (scenario analysis, prediction...) should always be in direct correspondence to the focus of the selected performance metrics. In other words, the lack of identifiability does not make the model structure useless for operation, but it limits the range of the application to the specific aim it was evaluated (calibrated) for, which is represented by the choice of performance metric.

In a next step, the optimal performance of the 24 model structures was compared to assess if a parameter optimization is able to make differentiation on the four defined model structural decisions (**Objective A.2**). The performance and resulting hydrograph after optimization to a specific metric were very similar amongst the members of the ensemble. The effect in variation of individual model structure components (one process at a time) could not be effectively assessed based on the performance metric in the case of the VHM model. For the four model decisions defined based on VHM, parameter optimization was not useful to differentiate model structures within the flexible model environment. At the

same time, the parameter values corresponding to the optimal performance varied largely amongst the different ensembles, notwithstanding their common task within the model structure.

The latter is probably due to the insufficient parameter identifiability of the individual members of the used ensemble in this study, leading to the inability to distinguish them based on the used performance metrics. In other words, the parameter identifiability of the individual model structures is important to compare individual model decisions based on performance. For the chosen set of model decisions, the lack of parameter identifiability hampered the execution of the diagnostic approach.

Qualitative assessment

As the different model representations cannot be distinguished based on their performance, a new graphical method was presented that could still provide useful information, even though identifiability lacks. The method fits in the metric oriented approach, as it can be applied to a variety of user-defined metrics.

The qualitative and visual technique builds on the characteristics of a SA and the methods developed in Part II. Similar to SA techniques that quantify sensitivity based on a one at a time adaptation of a parameter value, the technique evaluates the changes in function of a single model adaptation. By an interpretation of the shift in parameter influence induced by a single model adaptation, the relevance of the related model component towards the used performance metric can be assessed.

Applied to the case study of the Nete presented in chapter 6 it could be concluded from the analysis of the selected performance metrics that the choice of a non-linear storage component is recommended, whereas the usage of an inter flow component and the antecedent rain concept are mainly necessary to represent the low flow conditions. Besides, routing can be simplified.

The generalisation of the concept of parameter sensitivity analysis towards model component sensitivity analysis is a useful concept. It enables the guidance towards a more identifiable model structure as intended by **Objective A.3**.

Limitations of the VHM approach

The VHM approach is unique in the way it provides a step-wise approach to setup the model and in the derivation of the parameters based on different transfor-

mations of the available flow, rainfall and evapotranspiration observations. The VHM approach uses different information sources (derived from the original flow time series) in the calibration and model building process, while keeping the model structure flexible. As such, it fits in the methodology of this dissertation and it provides an embedded multi-criteria evaluation of the model performance making the resulting parameter values more robust. However, at the same time, some shortcomings should be emphasised, to understand the limitations of the model and in order to take the next steps.

First, in the original contribution of Willems (2014), the authors do refer to a *parsimonious* modelling approach, which is something I do not agree with. The terminological indistinctness of the word parsimonious modelling was described in section 2.5.1, as well as the proposed handling of parameter identification as model property. Moreover, in case of the VHM approach, with a total number of parameters ranging from 9 till 15, it is rather contradictory to call the approach *parsimonious*, even though the calibration is done component based on different data-derived information streams. The model structures involved are all practically not identifiable, as illustrated by the similar performance achieved by them in chapter 7.

Secondly, the implementation of the VHM model is obscure with regards to the different fractions calculated and with how to keep the balance straight. The distinction between the mathematical and computational model cannot be made. The closure of the mass balances by making sure the sum of the fractions is one, is actually a model structure decision on itself. It seemed that the fractions as presented in Willems (2014) do need adaptation to make sure the mass balance stays correct which was not reported on.

The VHM approach acts as a valve distributing the incoming rainfall amongst the different components. By applying a splitting of the incoming rainfall, it is unique in this sense to most other lumped hydrological models. But from a perceptual view, rainfall directly contributing to the groundwater, does not really represent the catchment behaviour. In many cases, groundwater is raised by percolation from the soil moisture component. When modelling a system by describing the different processes and their exchanges to make a representation of it, the perception of the reality should be reflected in the model structure (hypothesis of the system). Even in the case that this conceptualisation would be correct in some cases, the VHM approach is too limited in structural degrees of freedom to reflect a variety of existing catchments with different underlying processes.

Furthermore, when checking the output of the fractions plots, the occurrence of discontinuities are rather contra-intuitive, considering the discontinuous behaviour

of the fractions when the antecedent rain concept (representing the effect of infiltration excess) is active (Figure 6.8). These discontinuities should be avoided, both from a system representation point of view, where these jumps are not expected, as well as for model calibration leading to difficulties in the convergence towards an optimal parameter set (Clark and Kavetski, 2010; Kavetski and Clark, 2010, 2011).

Nevertheless the comparable rationale with the diagnostic approach and the advantages of using different data sources within the model identification process, these limitations make the VHM model structure implementations incompatible with the diagnostic approach. It declares the importance of a solid computational model (implementation) and proper model architecture as a minimal requirement to make a diagnostic approach possible.

As such, it can be concluded that we should always be aiming for identifiable and distinguishable model structures within a general flexible environment for modelling lumped hydrological models, while keeping the numerical implementation and discontinuity handling correct. These conclusions were taken as initial requirements in the next part of the dissertation.

11.4.2 Diagnosing structural errors in lumped hydrological models

Based on the VHM application, we learned on the one hand that the application of the diagnostic approach is only feasible when the computational and mathematical model are clearly separated. On the other hand it revealed that an identifiability analysis is the main driver in the identification of model deficiencies, as it is a necessary condition to attribute model differences to specific process adaptations as proposed by Clark et al. (2015b).

In Part IV, these two issues were tackled in order to fit the construction and evaluation of lumped hydrological models in the diagnostic approach proposed.

Towards reproducibility in hydrological modelling

Existing environments that support flexibility in the model building process for lumped hydrological models do mostly not comply with the requirements for a multiple hypotheses approach. Actually, both in the case of fixed model structures as well as flexible environments, the lack of transparency in the implementation and the inappropriate implementation of the computational model are the main weaknesses.

The FUSE concept proposed by Clark et al. (2008) provides an answer to this problem by translating existing lumped hydrological model structures in a general ODE mathematical model, similar to the central type of model structures studied in this dissertation (Equation 2.1).

To improve the current sharing of lumped hydrological models, a further generalisation of the FUSE approach is proposed. It summarizes the model in a matrix representation that is independent from the software or programming language used. It adapts the existing Gujer matrix representations for (bio)chemical ODEs model structures to cope with lumped hydrological model structures.

Specific attention is given to the translation of the NAM and PDM model into a set of ODEs. Both models are currently used in the operational water management in Flanders as part of a flood prediction system. It can be concluded that the ODE representations are not entirely the same as simulations of the original simulation platforms. However, the translation into a set of ODEs unifies both models and places these two specific model structures in a much wider framework of alternative model representations.

In Table 11.1 the matrix representation is verified towards the requirements for a multiple hypotheses approach. For each of the requirements, the properties of the matrix representation comply. Hence, when communicated together with the applied solver implementation (computational model), the diagnostic approach is supported.

Table 11.1: Assessment of the proposed matrix representation for lumped hydrological models with respect to the requirements of the multiple hypotheses approach.

requirements multiple hypotheses approach	properties matrix representation
alternative process descriptions	choice of constitutive functions
alternative interconnections and construction options	choice of the reservoir configuration
separation between mathematical and computation model	solver independent description of the model structure
accessible and transparent	open communication of the chosen model structure

A **standardised way of communicating about model structures supports reproducibility** in the application of lumped hydrological models. As intended by **Objective S.2**, the communication about hydrological model structures is made explicit and at the same time compliant tot he requirements of the multiple hypotheses approach. It removes the obscurity of existing implementations and cures us from the fetish towards model name acronyms.

Time-variant model evaluation

In the last part, different aspects of the dissertation were brought together by performing a model evaluation of the NAM and PDM lumped hydrological model structures.

The DYNIA method is of particular interest for model structure evaluation as it provides insight in the parameter identifiability as a function of time. Due to the importance of the uncertainty of the discharge when derived from a rating curve analysis, it was decided to incorporate this information in the performance metric construction as limits of acceptability. In a final step, the GLUE approach was used to assess and compare the effect of a chosen threshold on the variability of the model output for both models.

The application provided useful information about both model structures. Whereas it is known that the groundwater representation is essential to capture the seasonal dynamics of the Nete catchment, both models tackle it differently but both approaches show shortcomings. Furthermore, the probability based soil storage representation of the PDM model outperformed the NAM structure. Still, it is important to understand that the derived information about the model structural behaviour of both NAM and PDM are function of the observed time series used and the characteristics of the Nete catchment and should not be generalised.

Hence, the main result is that a time-variant evaluation (in this case DYNIA) provides guidance towards both model optimization and identification, as was intended by **Objective A.4**.

Periods of high influence of the parameters can be identified using the graphical output of the DYNIA method. These periods provide the best chance of estimating parameters more reliable and should be used in the aggregation or performance metrics.

When selecting a set of performance metrics, it is not always straightforward to identify a set of complementary metrics, each focusing on a different aspect of the model performance (Gupta et al., 1998). The outcome of the DYNIA method

provides information about useful aggregations and metrics. For example, in the application of the Nete catchment, the use of a total seasonal volume could support the model optimization for practical applications.

Summarized, the DYNIA method or, more general, **a time-variant approach of model evaluation**, provides a general scan of the model behaviour which helps to identify deficiencies, as an x-ray scan enables a medical doctor to make a further diagnosis about the injuries.

As such, time variant parameter identifiability provides a promising research perspective that should be further developed and made (publicly) available to a wider audience. Initiatives such as the temporal performance evaluation by Reusser et al. (2009), which is directly accompanied with a supporting R, package should be supported.

CHAPTER 12

Perspectives

Starting from current limitations of environmental modelling practices with respect to model identification and evaluation, an alternative **diagnostic approach** has been proposed and applied in this dissertation. Based on a flexible implementation of the mathematical and computational model on the one hand and an improved integration of model evaluation tools and performance metrics on the other hand, a step towards the assessment of individual model decisions is taken. However, further elaboration is needed to make this approach work in an operational setting and the approach is prone to discussion.

A drawback of proposing multiple working hypotheses as different model representations was already noticed by Chamberlin (1965). It is far easier for students, practitioners and stakeholders to accept a single interpretation (model) as a representative to apply than to recognize the several possibilities and putting them into a learning framework.

Furthermore, one could argue that the existence of the huge variety of environmental scientific disciplines using modelling approaches and relying on its own modelling traditions is an organically grown response to the need of flexibility in the modelling approach. This could be considered as a good thing, since each community develops a tailor-made approach. However, the sprawl of semantic discussions and obscurity in terminology within and in between them illustrates the consequence of this situation. The reality is that current fragmentation in between disciplines lead to redundancy and inferior practices. Counteracting this situation is not evident and some level of redundancy will always be existing.

However, each scientist should have the continuous ambition of keeping inferior practices and redundancy to a minimum. Only by accepting the modularity in

modelling and seeking a model fit for purpose, within the uniqueness of each place and application, scientific research can make progress. Confronting the features of a large set of (tailor-made) model structures and the properties of the corresponding applications (system characteristics, data, research question,...) could eventually lead towards more unified theories. It provides the opportunity to recognize patterns on a larger level. Current practices of tuning existing monolithic models towards each new application will only result in more tweaking of parameters to make model outputs fit with the observations. This sense of positivism without identification of the deficiencies of the model structure itself does not support scientific knowledge on the long run.

It is important to understand that the flexible approach is not a statement against detailed model descriptions. Under the assumption of sufficient data and when it supports the research question, it would be very conservative to be against a more detailed description. The main requirement for an identifiable model is the proper balance between data availability and model complexity. Flexibility in model development is the key to find this balance considering the uniqueness of each model study. Each component in the model has a specific function in the conceptual representation and should be identifiable as such.

Real world observations are needed to confirm the proper functioning of the considered components. This does not mean that all parameters need to be identifiable during the whole simulation. Processes are active during different time periods (cfr. difference between wet and dry weather conditions) which should be represented by changes in the sensitivity of model components as well. When the real-world observations do not represent the processes included in the model representation, identifiability will be hampered.

This is why time-variant methods for sensitivity and identifiability are essential in the model evaluation process. It enables the modeller to evaluate if components are representing the real-world processes as intended and to assess the consistency of the representation in function of time. This is also why predictions outside the range of data characteristics tested with (calibration and validation), will always be prone to uncertainty.

This dissertation supports future modellers in understanding the modelling terminology, putting existing methods and models in the right perspective and as such, to improve their model evaluation and application capabilities. An important step is the availability of a modular and transparent set of tools (which can be adapted).

Still, drawbacks and failures are present in the implementations. The original design goals of the implementations do not comply any more with current best practices of scientific computing (Wilson et al., 2014). The awareness towards the underlying architecture (code, code structure, code documentation...) of the implementations has been a gradual process in line with a growing concern in the broader scientific world (Prabhu et al., 2011).

In this final chapter, the further development of a diagnostic approach is framed in a broader scientific perspective. Starting from a Mea Culpa in the practical execution of the diagnostic approach, a further perspective is provided.

12.1 Mea culpa

A major failure that can be addressed is the trap that many researchers seem to fall into: the creation of yet another set of packages by a single contributor, trying to capture a range of functionalities, which seem to be limiting after all. During the execution of the dissertation, the illusion of yet another package for model evaluation methods was considered. It is doomed to again be used by only a small community of believers and die silently on the graveyard of good intentions. As such, the proposed solution is actually the engine itself of the scattered development. The acronym-fetish towards model structures has been converted to a fetish of acronyms for new packages.

Whereas the intention of making the methods and applications accessible to others can be encouraged, simply the fact that the implementations are shared and open does not make it superior or better.

It is clear that more transparency in modelling applications and the availability of tools for model evaluation can counteract the conservatism discussed in the beginning of the dissertation. The question is on how we can make progress as a modelling community, taking into account the need for a reproducible scientific practice, the dissemination of good practices and the reduction of redundant work by individuals. In the next sections, I will elaborate on a possible way forward, based on the experiences gained during the execution of the work.

12.2 Modularity as scientific good practice

Monolithic model implementations were identified as a major drawback for environmental modelling. They hamper the evaluation of individual model processes and they do not align with the need of adaptation towards changing conditions. Flexible model environments that comply to the requirements of a multiple hypotheses approach overcome this drawback. Flexibility is practically provided by a modular implementation of corresponding components. The latter is not new in integrated modelling (Voinov and Shugart, 2013) and is also reflected in the numerous environments for modular model development (section 2.5.2).

This trend towards modularity is also seen in computer software design. It aims to break monolithic software down into many separate components (microservices) which operate together as a whole. When different components provide a service to other components over a network using a communication protocol, this is referred to as a Service-Oriented Architecture (SOA). Following quote by Newman (2015) about the advantages of microservices can be directly transferred with the flexibility arguments for modelling as well:

> With a system composed of multiple, collaborating services, we can decide to use different technologies inside each one. This allows us to pick the right tool for each job, rather than having to select a more standardized, one-size-fits-all approach that often ends up being the lowest common denominator... With microservices, we are also able to adopt technology more quickly, and understand how new advancements may help us.

By making the creation of independent and reusable functionalities the goal of any kind of implementation, re-usage is possible, redundancy (*copy-paste* behaviour) is reduced and automation is supported.

The necessity of **modularity** is not only a model building requirement, but should be extended towards a more **general good scientific practice** (Wilson et al., 2014). When created as independent functionalities, entities can interact with one another and implementations easier shared. It can be further extended into reproducible workflows for which each of the steps can be interchanged when needed.

Scripting languages such as R and Python are gaining popularity as a fast and reliable way to modular and flexible development (Vitolo et al., 2015). Basically, every function created in Python or R is already an independent functionality that

can be integrated in a wider workflow (pipeline). It directly counteracts current practices in Graphical User Interface (GUI) based spreadsheet software that lead to unreproducible workflows, lack version control and limit automation.

Some of the developed functionalities will be useful for a wider audience. As proposed by Buytaert et al. (2008), commonly used routines and processes can then be implemented as generic software libraries in a low-level language such as C or Fortran and reused in virtually every environment. Actually, for some numerical solvers, this is already the case and these libraries are used in commercial applications as well (Hindmarsh, 1983).

The question is how this process can be managed. The risk is that it results in a variety of similar packages doing similar things and all developed by a single developer (mea culpa). Some redundancy will always exist and competition can also accelerate new developments. However, the key to a more successful development is the collaboration across the boundaries of scientific disciplines, which will be explored in the next section.

12.3 Towards community based collaboration

Environmental modelling is an interdisciplinary field, relying on computational and mathematical knowledge to study the natural environment. However environmental scientists are not trained in all aspects of computation and math and need to rely on external knowledge. Collaboration is essential, but not always feasible and is directly dependent from the network working in, making it sometimes ad hoc. Collaboration should be feasible on a much larger scale.

The current success of open-source software development illustrates the potential of a collaborative development across different disciplines. Python Pandas (McKinney, 2010) is supported by over 500 contributors and spans a wide range of disciplines with both industrial and academic backgrounds. The environmental modelling community can learn a lot from open-source development, where functionalities are available as packages and libraries which can be forked, adapted and extended.

The main reason of the successful collaboration is the technological advancement, making global communication possible (cfr. the digital revolution is able to support *commons* on a larger scale). **Online curated code repositories** such as Github, Gitlab and Bitbucket, provide a platform for online collaboration. Code can be revised, features can be discussed and the history of the code development

is tracked by the revision control system. It is a transparent system, making it possible for anyone to cooperate.

These environments can support the collaboration across the boundaries of scientific disciplines and we should take advantage of this opportunity. In short, **we should build our tools on the shoulders of giants**, i.e. the open source communities active world-wide and continuously developing improved tools in their field of expertise.

Consider following example. The increasing popularity of Bayesian applications within the hydrological modelling community appears to result in a narrow range of applied methods (mainly BATEA and DREAM), both dependent on an MCMC sampler. The sampler scheme itself is nested within the code. Though, the field of Bayesian computing is fast evolving an improved sampling strategies are constantly developed, which would be interesting to test as well. A proper decoupling of the sampling strategy would enable to anticipate to the continuous development achieved in mathematical and statistical research, made available by open source libraries focusing on MCMC sampling (Davidson-Pilon, 2015). At the same time, more fundamental research communities are able to make their developments available to a wider audience by contributing to these libraries.

The aim is to make sure that each scientific community can focus on their specific specialisation, respecting the qualification of other communities and building on each other strengths. By doing so, we can continuously rely on these communities provide the theoretical and technical foundations that we need to build our domain specific technology on. The metric oriented approach fits in this prospect, putting the focus on the domain knowledge, while relying on external knowledge for sampling and optimization.

Hence, this is an advocacy towards a more **collaborative code development**, where code revision within the community is a continuous process, just as it is with publications. It enables a continuous development cycle, where more revision by more partners can lead to an accelerated development and more scrutiny. It counteracts the regularly seen central-development approach, where a single group is 'providing' their methods as a black box towards a wider community (Kuczera et al., 2006; Pianosi et al., 2015; Vrugt, 2015), which is not transparent at all.

The current success of open source scripting languages, such as R and Python, do already support collaboration by a continuously growing group of users. An increasing trend in the usage of open source developments for research purposes is already observed. However, modular code implementation, code sharing and collaborative development as a scientific good practice is not yet embedded in

current environmental modelling practices. In the next section, the perspective of an open science policy is put forward as an engine for collaboration and accelerated progress.

12.4 Open science as an engine for collaboration

Access to the implementation is important for a fundamental aspect of scientific practice. The entire idea of scientific peer-review is based on the ability to reproduce the results. Reproducibility of computational methods is only possible when the entire implementation is available (Peng, 2011). However, the publishing and sharing of code is still lagging behind (Buytaert et al., 2008). Focus is currently still on the publication itself, which is only the minimal level on the entire spectrum of reproducibility (Figure 12.1).

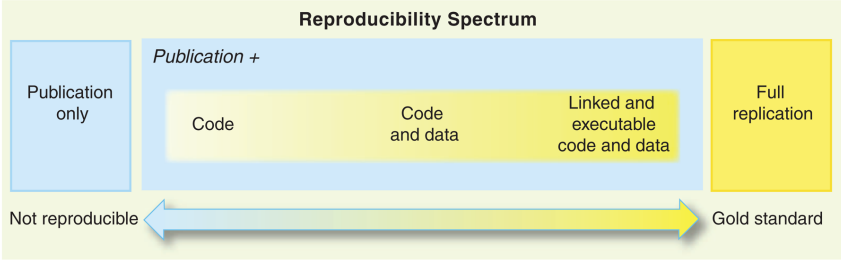


Figure 12.1: The spectrum of reproducibility. Current common practice of scientific publication peer review only supports a very minimalistic level of reproducibility. The necessity of sharing both code and data is essential to enable replication of scientific studies (Peng, 2011)

When new methodologies are proposed in literature, but the implementation is not available, it hinders the execution by the peer researchers and limits scientific progress. Scientific investigation should be open and transparent to ensure direct reproducibility and repeatability. It requires a change in mindset for current scientific practice, but provides many opportunities as well. As scientists, we should not be ignorant to this necessity.

In the following sections, we will discuss this from the perspective of respectively the scientific practice, the scientific education and the private sector.

12.4.1 An open scientific practice

Current scientific progress based on a peer review process of papers does not provide the incentives to researchers to share their implementation. Many environmental scientists are reserved about sharing their code. Actually, most scientists understand the importance of scientific computing (Prabhu et al., 2011), but do regularly not know how reliable the software they use actually is (Wilson et al., 2014).

In other words, development of software tools are not regarded as a scientific contribution and academic environments do not reward tool builders (Prabhu et al., 2011). The current focus on the achievement of publications results in reduced attention towards the implementation itself, which is however the central part of environmental modelling. Proper scientific attribution for software citation is currently lacking.

The advantages of publishing source code in an organized and proper manner are however evident, similar to the benefits of data sharing (Roche et al., 2015). It allows other scientists to reproduce prior work and compare new contributions on an equal footing. Researchers do not have to spend time rewriting the same pieces of code. It enables revision of code by other scientists, guarding against the bugs everyone inevitably makes and improve readability (Wilson et al., 2014). Similar to open data initiatives, the public sector should take a leading role by demanding open access by default. Sharing accelerates scientific discoveries and can save taxpayers' money by avoiding unnecessary duplication.

Hence, code implementation is as a fully fledged part of the experimental apparatus and should be built, checked, and used as carefully as any physical apparatus (Wilson et al., 2014).

To facilitate this process, code revision should become an essential part of the scientific investigation. However, a rigorous review of a computational method implementation will typically take longer than that of a more traditional paper (Editorial, 2015). Hence, researchers should get explicitly rewarded for their contribution to code development and revision. This means attribution for the creation of new code, but even more important, **scientific attribution for the revision and improvement of existing code**.

Scientists should not continuously create new packages, but collaborate on the development of functionalities, using the current technological features provided by online curated code repositories. The latter is the best antidote against the fetish of acronyms and a crucial incentive for collaborative development. Furthermore, it

would support continuity in scientific development across the borders of individual projects and dissertations.

Furthermore, classic scientific communication based on journal papers is not the appropriate communication channel for code collaboration and collective development as it is proposed here. It requires a fast communication medium where technical adaptations can be directly discussed by the community. We should explicitly discuss and express the collaboration on code development, taking it away from the corridor discussions at conferences to the plenary sessions.

12.4.2 Preparing future environmental modellers

Recent studies have found that scientists typically spend 30% or more of their time developing software, whereas 90% or more of them are primarily self-taught programmers (Wilson et al., 2014). Current environmental scientists lack exposure to basic software development practices such as writing maintainable code, using version control and issue trackers, code reviews, unit testing, and task automation.

These skills are essential to make an open and reproducible scientific practice successful. At the same time, environmental modellers should not all be trained computer engineers. An equilibrium needs to be searched for, which requires changing the features of the software systems scientists use on the one hand and getting researchers to work with systems supporting reproducibility on the other hand (Peng, 2011; Shou et al., 2015).

The former is a transition currently going on. For example, OpenRefine has a history that can be exported along with the data and imported back in to OpenRefine to reproduce the analysis (Verborgh and De Wilde, 2013). The latter is shifting as well. Lab skills for research computing are getting increased importance in the curriculum of environmental education. The growing success of international workshops such as software carpentry illustrate the awareness (Wilson et al., 2014). Emerging technological developments, such as the Jupyter notebook (Shen, 2014), provide an interactive computing environment that directly facilitate the reproducibility of the executed work.

To support reproducibility, **the competence of writing re-usable functions that are small enough to test and reuse**, should be central in the education of environmental scientists. Just making code available is not enough, the way in which it is done, is as important as the delivered code itself.

Furthermore, the **contribution towards open source code projects should be part of the scientific curriculum of every environmental modeller**. Bug fixing and coding new features would be too ambitious at the start. However, writing documentation, participating to discussions about issues, diagnosing bugs, writing tests and creating examples are definitely evenly useful. By doing so, the essential competences for a reproducible and collaborative scientific practice are acquired, while the continuity in development is assured.

12.4.3 A business model for open science

Closed source (proprietary) modelling software still constitutes an important part in the scientific literature. Reporting scientific analysis based on closed source model environments hampers reproducibility and is unfair to scientists without the access to the necessary licenses. Closed source environments can only be changed by their owners, who may not perceive reproducibility as a high priority (Peng, 2011).

At the same time, (closed source) software development also facilitates the development and distribution towards practitioners of good modelling practices. It provides the essential software backbone and enables the computational optimization. A competitive market will stimulate the innovation and accelerate incorporation of new technologies.

We should strive to combine the strengths of both worlds. A distinction needs to be made about the modules required for a scientific investigation (model components, algorithms for model evaluation...) and the elements of the GUI that facilitate the user experience.

The former elements need to be embedded in a scientific reproducible practice with accessibility of the code, whereas the latter provides the opportunity for software development companies to differentiate themselves from both competitors and a script-based approach.

Actually, this contributes to the idea of collaboration across community boundaries (section 12.3). The elements that are fundamental part of the scientific research are developed as a collaborative effort between both research institutions and software companies. It provides a solid layer supported by scientific research and can be cited as such. At the same time, the implementations are accessible to anyone who wants to create a product or application from it, facilitating the user experience.

Users who do not have the competences to work with the implementations directly, will rely on these applications. Still, this does not limit the scientific reproducibility. The essential building blocks for model construction and evaluation directly rely on the publicly accessible code and can be referenced as such. Product developers are able to close the parts of the code that contribute to user experience, but are obligated to keep the core functionalities of the mathematical and computational model as well as the model evaluation methods accessible. This is made possible by the modular approach of implementation (section 12.2) and by **proper licensing of the different components to determine responsibility of the users** (Roche et al., 2015). By providing an open source license, the conditions on how to use and collaborate on the code are stipulated. Adding no license at all means that default copyright laws apply and that nobody else may reproduce, distribute, or create derivative works from the code[1]. An open-source license allows reuse of your code while retaining copyright. Hence, they provide the necessary terms on which collaboration on a community level can be expressed and can actually counteract misuse.

This perspective is actually a translation of the current open source software business models, illustrating the huge potential of this approach. Indirectly this actually already happens, since we constantly use functionalities written in some language and provided by someone. By making this explicit, environmental modelling would become much more democratic and fair on a global scale.

This does by no means threaten the service oriented business model of consultancy companies active in the environmental sector. On the contrary, it can potentially diminish the false concurrency of universities and other public institutes, since scientific developments and tools are accessible and directly available. As such, collaborations between public and private partners are not a necessity in order to have code access, but a collaboration of specific service and knowledge. It also opens perspectives to a more competitive tender application, since all applicants can start from a common accessibility to the fundamental implementations. Hence, creativity and excellence will be the key drivers.

12.5 Need for standardisation

Open science supports collaboration, since it provides the ability to integrate the work of others. The main obstacle to take is the communication in between the different actors, otherwise the incoherence in terminology will hinder progress (sec-

[1] http://choosealicense.com/no-license/

tion 2.4.1). The ability to interconnect model components and methodologies (connectability) should be regarded as important as the accessibility of the implementation itself (Kraft, 2012; Le Phong et al., 2015). **Interoperability of building-blocks** is a major source of concern which **can be enabled by defining standards** (Vitolo et al., 2015).

To ensure consistency among concepts belonging to similar scientific disciplines and across disciplines, standardization of definitions, data and formats is continuously needed. The standards managed by the OGC (Open Geospatial Consortium) for geospatial data, such as the WaterML 2.0 for water observations data and the Open Modelling Interface (OpenMI) for the exchange of data between process simulation models, are examples of existing standards relevant for the water community.
The internet provides the most universal communication platform currently available, so compliance with the open standards provided by The World Wide Web Consortium (W3C) is essential to exploit the abilities of the web. Hence, standardised web services provide the best chance for the sharing of information in between components (and communities). It enables standardized data exchange which can be used to chain different functionalities into complex workflows (Vitolo et al., 2015).

12.6 Closure: A perspective for the implementations

This chapter started with the awareness about the limitations of the translation of the diagnostic approach towards a practical working scheme. The perspective of an open and reproducible scientific practice is a main driver to overcome the conservatism in environmental modelling in direct support of the diagnostic approach. It guards against protectionism and it inherently provides flexibility in both model construction as well as evaluation.

Part of the work of this dissertation has been made available online. So, what is the perspective of the developed packages?

The integration of the pystran Python Package 4 with comparable initiatives (Houska et al., 2015; Usher et al., 2015) is a major perspective to ensure the continuity of the implementations and the work. Furthermore, the package should be dismantled into two major parts to better support the metric oriented approach.

The first part should be completely oriented on the creation of performance metrics, further extending the existing functions to develop metrics as well as more theoretical descriptions (e.g. likelihood functions). This can considerably extend the exploration and diagnosis phase of model structures and overcome conservative model evaluation practices. A clear selection on the interactions with existing major packages is a crucial element to ensure good practices in terms of optimization and sampling.

The other part should focus on the further development of methods for sensitivity and identifiability analysis with a particular focus on time-variant methods. The main design goal should be the ability to recycle simulations as efficient as possible among different algorithms to maximize the extracted information (section 5.10.2).

The development and integration of machine learning techniques within the scope of the sklearn package in Python could serve as a blue print on how a set of algorithms can be collected within a rigid framework (Pedregosa et al., 2011; Buitinck et al., 2013). The library is developed by an international community, with a focus on maintainability by using strict quality guidelines about code consistency and unit-test coverage.

The hydropy Python Package 1 represents another type of development which has only been shortly mentioned in the dissertation. It provides a practical support in the calculation of aggregated metrics. It already relies on a *giant* to ensure the base functionalities and just adds a small layer of domain knowledge on top of it. Ensuring compatibility with the Pandas package is the main perspective, while gradually adding alternative domain-specific methods. Further development is currently conducted within the own research unit, adding additional classes for handling time series originating from a lab-based environment. External collaborators are invited to contribute to the code.

Other implementations are available on Github[2] and can be used and further improved by other users. Furthermore, the flowchart to provide guidance on the selection of a sensitivity analysis method is available. Github provides an appropriate online environment to collaboratively discuss, adapt and improve it in the future.

A similar exercise could be useful for the standardised matrix representation for lumped hydrological models. Making the further development an open and transparent discussion could potentially provide it the leverage it needs to be generally accepted. Another useful perspective is the extension towards a generic

[2]https://github.com/stijnvanhoey

model description and implementation for spatial explicit (distributed) hydrological modelling according to the requirements of the diagnostic approach. The most known distributed hydrological model, providing a range of process descriptions, is MIKE-SHE (Refsgaard and Storm, 1995). It provides an OpenMI interface (Moore and Tindall, 2005) for coupling with other models, but fails at the request for code accessibility. Both the model building approaches of Kraft (2012) and Clark et al. (2015b,c) are open access, using a set of conservation equations and are providing flexibility in the structural configuration, while keeping the mathematical and computational model separated. They comply to the requirements and should be further supported by the hydrological modelling community. In combination with an extension of the matrix representation towards PDEs, reproducibility would be supported on a distributed level as well.

Still, lumped hydrological models should be treated as a set of ODEs and communicated as such, supported by the standardized matrix representation. This also means that code contributions should go to modelling environments supporting the implementation of any set of ODEs, such as the development of the pyideas package in Python (Van Daele et al., 2015c) or the deSolve package in R (Soetaert and Petzoldt, 2010b).

PART VI

APPENDICES

APPENDIX A

Additional figures for DYNIA application

In this appendix, the DYNIA plots are given for the remainder of the parameters not provided in the main text.

A.1 PDM model

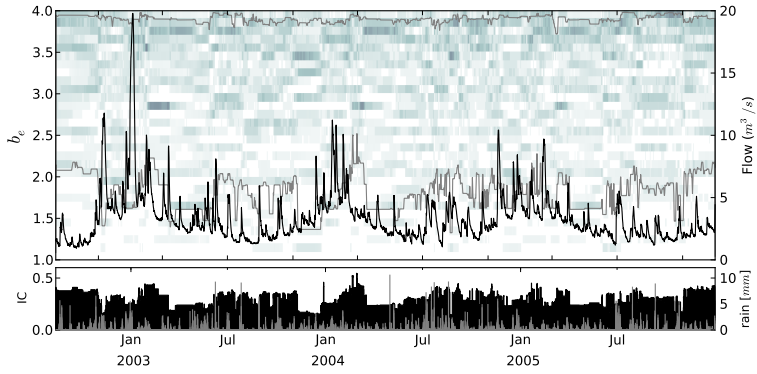


Figure A.1: Results of the DYNIA procedure for parameter b_e (PDM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

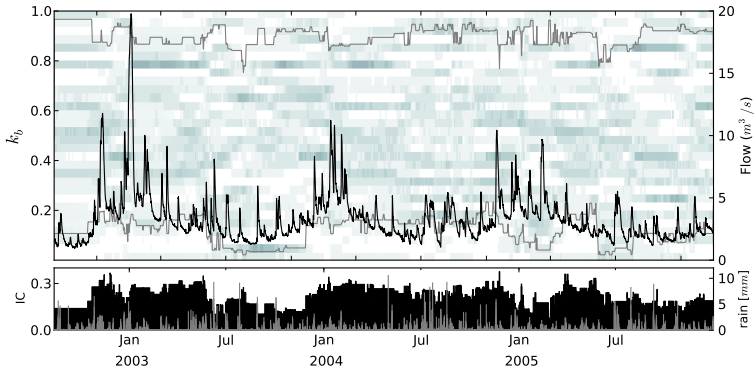


Figure A.2: Results of the DYNIA procedure for parameter k_b (PDM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

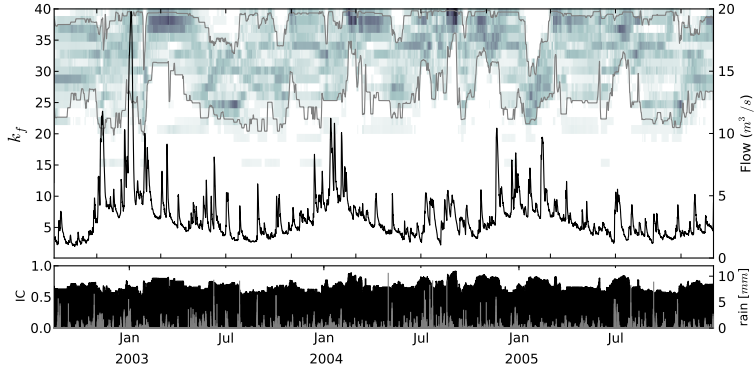


Figure A.3: Results of the DYNIA procedure for parameter k_f (PDM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

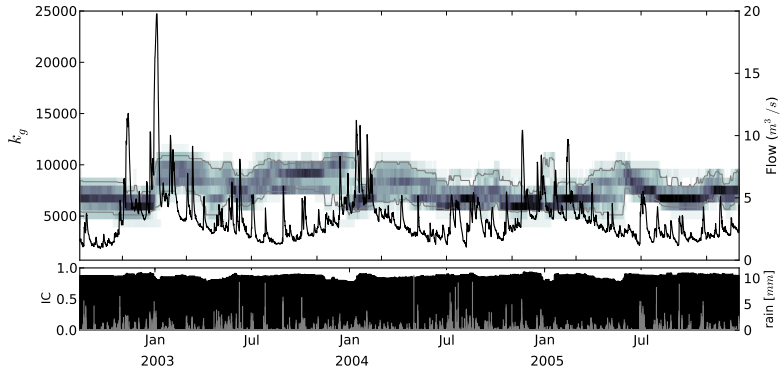


Figure A.4: Results of the DYNIA procedure for parameter k_{g} (PDM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

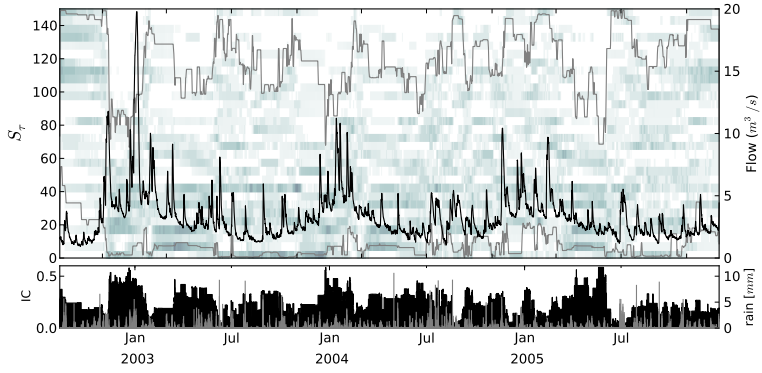


Figure A.5: Results of the DYNIA procedure for parameter S_τ (PDM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

A.2 NAM model

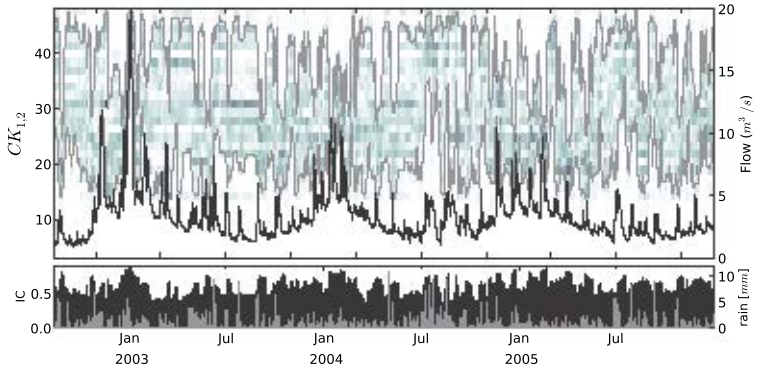


Figure A.6: Results of the DYNIA procedure for parameter $CK_{1,2}$ (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

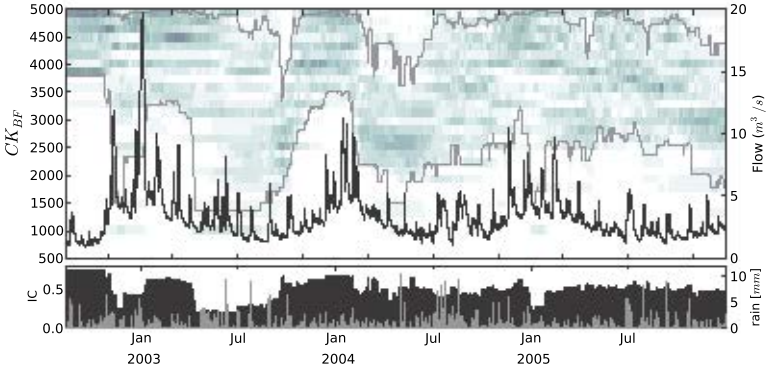


Figure A.7: Results of the DYNIA procedure for parameter CK_{BF} (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

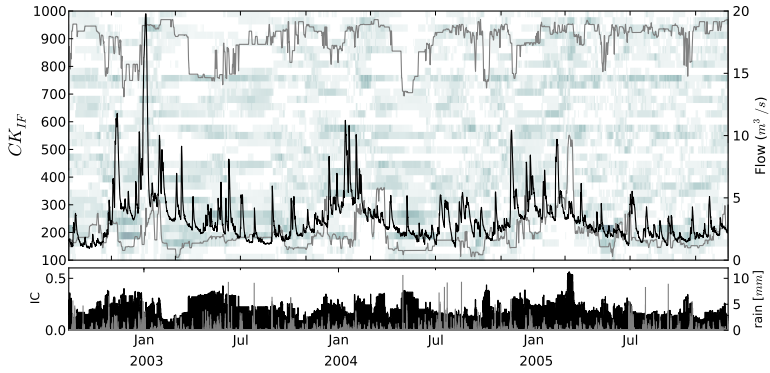


Figure A.8: Results of the DYNIA procedure for parameter CK_{IF} (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

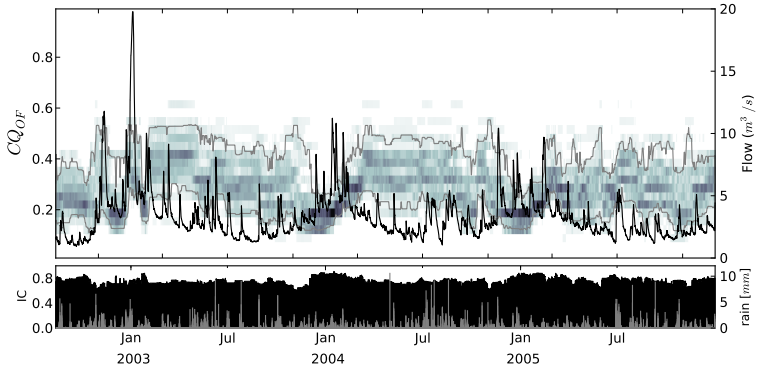


Figure A.9: Results of the DYNIA procedure for parameter CQ_{OF} (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

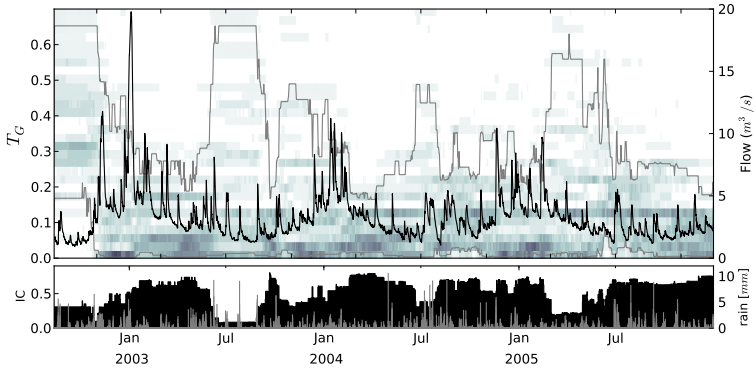


Figure A.10: Results of the DYNIA procedure for parameter T_{G} (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

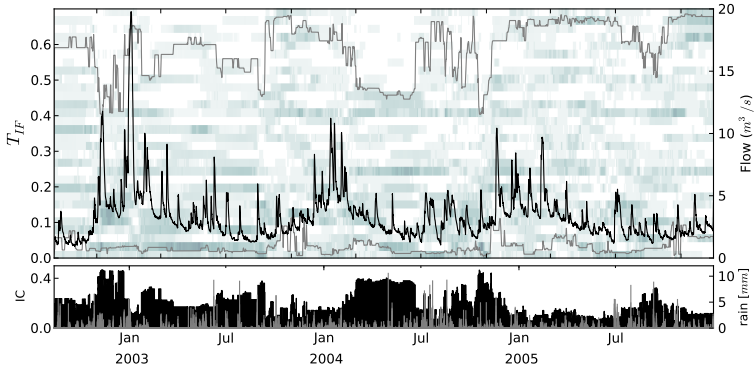


Figure A.11: Results of the DYNIA procedure for parameter T_{IF} (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

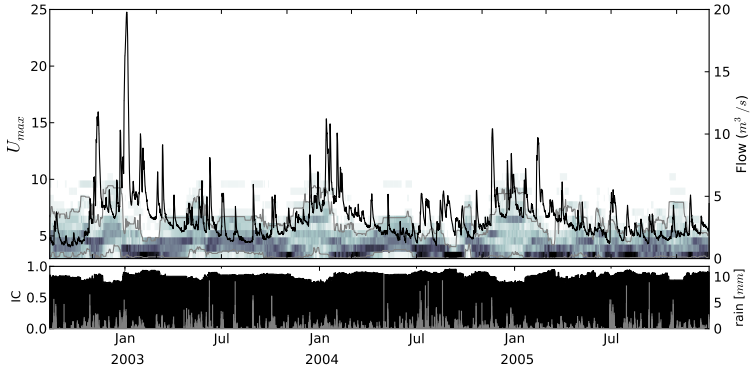


Figure A.12: Results of the DYNIA procedure for parameter $U_{\max}$ (NAM model) applied to the behavioural model simulations for the calibration period (see Figure 10.8 for explanation).

References

Abbaspour, K. C. (2005). Calibration of hydrologic models : When is a model calibrated? In Zerger, A. and Argent, R., editors, *MODSIM05, International congress on Modelling and simulation advances and applications for management and decision making (2005)*, pages 2449–2455, Melbourne, Australia.

Abebe, N. A., Ogden, F. L., and Pradhan, N. R. (2010). Sensitivity and uncertainty analysis of the conceptual HBV rainfallrunoff model: Implications for parameter estimation. *Journal of Hydrology*, 389(3-4):301–310.

Abramowitz, G. (2010). Model independence in multi-model ensemble prediction. *Australian Meteorological and Oceanographic Journal*, 59:3–6.

Anderton, S. P., Latron, J., White, S. M., Llorens, P., Gallart, F., Salvany, C., and O'Connell, P. E. (2002). Internal evaluation of a physically-based distributed model using data from a Mediterranean mountain catchment. *Hydrology and Earth System Sciences*, 6(1):67–83.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L. (2012). All that glitters is not gold: the case of calibrating hydrological models. *Hydrological Processes*, 26(14):2206–2210.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A. (2009). Hess Opinions: Crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences*, 13(10):1757–1764.

Andréassian, V., Perrin, C., Parent, E., and Bárdossy, A. (2010). The court of miracles of hydrology: can failure stories contribute to hydrological science? *Hydrological Sciences Journal*, 55(6):849–856.

Andrews, F. T., Croke, B. F. W., and Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. *Environmental Modelling & Software*, 26(10):1171–1185.

Argent, R., Brown, A., Cetin, L., Davis, G., Farthing, B., Fowler, K., Freebairn, A., Grayson, R. B., Jordan, P., Moodie, K., Murray, N., Perraud, J.-M., Podger, G. M., Rahman, J. M., and Waters, D. (2008). WaterCAST Manual.

Argent, R. M. (2004). Concepts, methods and applications in environmental model integration. *Environmental Modelling & Software*, 19(3):217.

Argent, R. M. (2005). A case study of environmental modelling and simulation using transplantable components. *Environmental Modelling & Software*, 20:1514–1523.

Argent, R. M., Voinov, A., Maxwell, T., Cuddy, S. M., Rahman, J. M., Seaton, S. P., Vertessy, R. A., and Braddock, R. D. (2006). Comparing modelling frameworks - A workshop approach. *Environmental Modelling & Software*, 21(7):895–910.

Arnaldos Orts, M., Amerlinck, Y., Rehman, U., Maere, T., Van Hoey, S., Naessens, W., and Nopens, I. (2015). From the affinity constant to the half-saturation index: understanding conventional modeling concepts in novel wastewater treatment processes. *Water Research*, 70:458–470.

Asprey, S. P. and Macchietto, S. (2000). Statistical tools for optimal dynamic model building. *Computers and Chemcial Engineering*, 24:1261–1267.

Bach, P. M., Rauch, W., Mikkelsen, P. S., McCarthy, D. T., and Deletic, A. (2014). A critical review of integrated urban water modelling - urban drainage and beyond. *Environmental Modelling & Software*, 54:88–107.

Bai, Y., Wagener, T., and Reed, P. M. (2009). A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, 24(8):901–916.

Beck, B. and Young, P. C. (1976). Systematic identification of DO-BOD model structure. *Journal of the Environmental Engineering Division*, 102(5):909–927.

Beck, M. B. (1986). The selection of structure in models of environmental systems. *The Statistician*, 35(2):151–161.

Benedetti, L., Batstone, D. J., De Baets, B., Nopens, I., and Vanrolleghem, P. A. (2012). Uncertainty analysis of WWTP control strategies made feasible. *Water Quality Research Journal of Canada*, 47(1):14–29.

Benedetti, L., Bixio, D., Claeys, F., and Vanrolleghem, P. A. (2008). Tools to support a model-based methodology for emission/immission and benefit/cost/risk analysis of wastewater systems that considers uncertainty. *Environmental Modelling & Software*, 23(8):1082–1091.

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.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, 4(2):203–213.

Beven, K. J. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, 5(1):1–12.

Beven, K. J. (2002). Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes*, 16(2):189–206.

Beven, K. J. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320:18–36.

Beven, K. J. (2008a). Comment on Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? by Jasper A. Vrugt, Cajo J. F. ter Braak, Hoshin V. Gupta and Bruce A. Robinson. *Stochastic Environmental Research and Risk Assessment*, 23(7):1059–1060.

Beven, K. J. (2008b). *Environmental modelling: An uncertain future?* Taylor & Francis.

Beven, K. J. (2012). *Rainfall-runoff modelling. The primer*. Wiley, 2nd edition.

Beven, K. J. and Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, 28(24):5897–5918.

Beven, K. J. and Binley, A. M. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298.

Beven, K. J. and Freer, J. E. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1-4):11–29.

Beven, K. J. and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24(1):43–69.

Beven, K. J., Smith, P. J., and Freer, J. E. (2007). Comment on 'Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology' by Pietro Mantovan and Ezio Todini. *Journal of Hydrology*, 338(3-4):315–318.

Beven, K. J., Smith, P. J., and Freer, J. E. (2008). So just why would a modeller choose to be incoherent? *Journal of Hydrology*, 354(1-4):15–32.

Beven, K. J., Smith, P. J., and Wood, A. T. A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, 15(10):3123–3133.

Beven, K. J. and Westerberg, I. K. (2011). On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, 25(10):1676–1680.

Blazkova, S. and Beven, K. J. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, 45(W00B16):1–12.

Boyle, D. P., Gupta, H. V., and Sorooshian, S. (2000). Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, 36(12):3663–3674.

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015–1030.

Buahin, C. A. and Horsburgh, J. S. (2015). Evaluating the simulation times and mass balance errors of component-based models: An application of OpenMI 2.0 to an urban stormwater system. *Environmental Modelling & Software*, 72:92–109.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A. (1973). A generalized streamflow simulation system - conceptual modeling for digital computers. Technical report, Joint Federal and State River Forecast Center, U.S. National Weather Service and California Department of Water Resources, Sacramento.

Buytaert, W., Reusser, D. E., Krause, S., and Renaud, J.-P. (2008). Why can't we do better than Topmodel? *Hydrological Processes*, 22(20):4175–4179.

Cabus, P. W. A. (2008). River flow prediction through rainfall - runoff modelling with a probability-distributed model (PDM) in Flanders, Belgium. *Agricultural Water Management*, 95(7):859–868.

Campolongo, F., Cariboni, J., and Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, 22(10):1509–1518.

Campolongo, F., Saltelli, A., and Cariboni, J. (2011). From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, 182(4):978–988.

Carstensen, J., Vanrolleghem, P. A., Rauch, W., and Reichert, P. (1997). Terminology and methodology in modelling for water quality management - a discussion starter. *Water Science & Technology*, 36(5):157–168.

Chamberlin, T. C. (1965). The method of multiple working hypotheses. *Science*, 148:754–759.

Changming, L., ZhongGen, W., Hongxing, Z., Lu, Z., and Xianfeng, W. (2008). Development of Hydro-Informatic modelling system and its application. *Science in China Series E: Technological Sciences*, 51(4):456–466.

Cierkens, K., Van Hoey, S., De Baets, B., Seuntjens, P., and Nopens, I. (2012). Influence of uncertainty analysis methods and subjective choices on prediction uncertainty for a respirometric case. In Seppelt, R., Voinov, A. A., Lange, S., and Bankamp, D., editors, *International Environmental Modelling and Software Society (iEMSs) 2012 International Congress on Environmental Modelling and Software. Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting*, Leipzig, Germany. International Environmental Modelling and Software Society (iEMSs).

Claeys, F. (2008). *A generic software framework for modelling and virtual experimentation with complex environmental systems.* Phd thesis, Ghent University.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S., Maxwell, R. M., Shen, C., Swenson, S. C., and Zeng, X. (2015a). Improving the representation of hydrologic processes in earth system models. *Water Resources Research*, 51(8):5929–5956.

Clark, M. P. and Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research*, 46(W10510):1–23.

Clark, M. P., Kavetski, D., and Fenicia, F. (2011a). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, 47(9):1–16.

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D., and Woods, R. A. (2011b). Hydrological field data from a modeller's perspective. Part 2: Process-based evaluation of model hypotheses. *Hydrological Processes*, 25(4):523–543.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M. (2015b). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, 51:2498–2514.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., Rasmussen, R. M., Tarboton, D. G., Mahat, V., Flerchinger, G. N., and Marks, D. G. (2015c). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, 51:2515–2542.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44(W00B02):1–14.

Crawford, N. H. and Linsley, R. K. (1966). Digital simulation in hydrology - the Stanford watershed simulation model IV, technical report No. 39. Technical report, Department of Civil Engineering, Stanford University.

Cullmann, J. and Wriedt, G. (2008). Joint application of event-based calibration and dynamic identifiability analysis in rainfall-runoff modelling: implications for model parametrisation. *Journal of Hydroinformatics*, 10(4):301–316.

Dams, J., Van Hoey, S., Seuntjens, P., and Nopens, I. (2014). Next-generation tools m.b.t. hydrometrie, hydrologie en hydraulica in het operationeel waterbeheer, fase 1: analyse, perceel 1: De Maarkebeek, standard evaluation hydrological models: G2G application. Technical report, Vlaamse Milieu Maatschappij.

David, O., Ascough, J. C., Lloyd, W., Green, T. R., Rojas, K. W., Leavesley, G. H., and Ahuja, L. R. (2013). A software engineering perspective on environmental modeling framework design: The Object Modeling System. *Environmental Modelling & Software*, 39:201–213.

Davidson-Pilon, C. (2015). *Bayesian methods for hackers: Probabilistic programming and bayesian inference*. Addison-Wesley Professional, 1st edition.

Dawson, C. W., Abrahart, R. J., and See, L. M. (2007). HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software*, 22(7):1034–1052.

De Pauw, D. J. W. (2005). *Optimal experimental design for calibration of bioprocess models: a validated software toolbox*. Phd thesis, Ghent University.

De Pauw, D. J. W., Steppe, K., and De Baets, B. (2008). Identifiability analysis and improvement of a tree water flow and storage model. *Mathematical Biosciences*, 211:314–332.

De Pauw, D. J. W. and Vanrolleghem, P. A. (2006). Practical aspects of sensitivity function approximation for dynamic models. *Mathematical and Computer Modelling of Dynamical Systems*, 12(5):395–414.

de Vos, N. J., Rientjes, T. H. M., and Gupta, H. V. (2010). Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrological Processes*, 24(20):2840–2850.

Decubber, S. (2014). *Linking the carbon biokinetics of activated sludge to the operational waste water treatment conditions*. Msc thesis, Ghent University.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag, New York.

DHI (2008). MIKE 11, A modelling system for rivers and channels, reference manual.

Di Baldassarre, G. and Montanari, a. (2009). Uncertainty in river discharge observations: a quantitative analysis. *Hydrology and Earth System Sciences Discussions*, 6(1):39–61.

Dochain, D. and Vanrolleghem, P. A. (2001). *Dynamical modelling and estimation in waste water treatment processes*. IWA Publishing.

Dochain, D., Vanrolleghem, P. A., and Van Daele, M. (1995). Structural identifiability of biokinetic models of activated sludge respiration. *Water Research*, 29(11):2571–2578.

Donckels, B. (2009). *Optimal experimental design to discriminate among rival dynamic mathematical models*. Phd thesis, Ghent University.

Duan, Q., Sorooshian, S., and Gupta, H. V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28(4):1015.

Duan, Q., Sorooshian, S., and Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, 158:265–284.

Editorial (2015). Reviewing computational methods. *Nature Methods*, 12(12):1099–1099.

Efstratiadis, A. and Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrological Sciences Journal*, 55(1):58–78.

Ekstrom, P. A. (2005). *A Simulation Toolbox for Sensitivity Analysis*. Msc thesis, Uppsala University.

Fall, A. and Fall, J. (2001). A domain-specific language for models of landscape dynamics. *Ecological Modelling*, 141(1-3):1–18.

Fenicia, F. (2008). *Understanding catchment behaviour through model concept improvement*. Phd thesis, Technische Universiteit Delft.

Fenicia, F., Kavetski, D., and Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, 47(11):1–13.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J. E. (2014). Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrological Processes*, 28(4):2451–2467.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G. (2008). Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research*, 44(W06419):1–13.

Filippi, J. B. and Bisgambiglia, P. (2004). JDEVS: an implementation of a DEVS based formal framework for environmental modelling. *Environmental Modelling & Software*, 19:261–274.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC Hammer. *PASP*, 125:306–312.

Foreman-Mackey, D., Price-Whelan, A., Ryan, G., Emily, Smith, M., Barbary, K., Hogg, D. W., and Brewer, B. J. (2014). triangle.py v0.1.1.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and GagneC. (2012). DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Re*, 13:2171–2175.

Freer, J. E., Beven, K. J., and Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, 32(7):2161–2173.

Frey, H. C. and Patil, S. R. (2002). Identification and review of sensitivity analysis methods. *Risk analysis*, 22(3):553–578.

Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C., and Di, Z. (2014). A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environmental Modelling & Software*, 51:269–285.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). *Bayesian data analysis, third edition*. Chapman and Hall/CRC, London, England, third edition.

Georgakakos, K. P., Seo, D.-J., Gupta, H. V., Schaake, J. C., and Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, 298(1-4):222–241.

Gernaey, K. V., Petersen, B., Nopens, I., Comeau, Y., and Vanrolleghem, P. A. (2002). Modeling aerobic carbon source degradation processes using titrimetric data and combined respirometric-titrimetric data: experimental data and model structure. *Biotechnology and Bioengineering*, 79(7):741–53.

Goodman, J. and Weare, J. (2010). Ensemble samplers with affine invariance. *Comunications in applied mathematics and computational science*, 5(1):65–80.

GPROMS (2015). Process systems enterprise.

Grady, C. P. L., Smets, B. F., and Barbeau, D. S. (1996). Variability in kinetic parameter estimates: A review of possible causes and a proposed terminology. *Water Research*, 30(3):742–748.

Gregersen, J. B., Gijsbers, P. J. A., and Westen, S. J. P. (2007). OpenMI: Open modelling interface. *Journal of Hydroinformatics*, 9(3):175–191.

Gujer, W. and Larsen, T. A. (1995). The implementation of biokinetics and conservation principles in ASIM. *Water Science and Technology*, 31(2):257–266.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48(W08301):1–16.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377:80–91.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763.

Gupta, H. V., Wagener, T., and Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, 22(18):3802–3813.

Hauduc, H., Neumann, M. B., Muschalla, D., Gamerith, V., Gillot, S., and Vanrolleghem, P. A. (2015). Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment. *Environmental Modelling & Software*, 68(3):196–204.

Helton, J. C., Johnson, J., Sallaberry, C., and Storlie, C. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91(10-11):1175–1209.

Henze, M., Gradyn, C., Gujer, W., Marais, G., and Matsuo, T. (1983). Activated sludge model no. 1. Technical report, IAWPRC task group on mathematical modelling for design and operation of biological waste water treatment processes. Technical report, IAWPRC, London, England.

Herman, J. D., Kollat, J. B., Reed, P. M., and Wagener, T. (2013a). Technical note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models. *Hydrology and Earth System Sciences*, 17(7):2893–2903.

Herman, J. D., Reed, P. M., and Wagener, T. (2013b). Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research*, 49(3):1400–1414.

Herron, N., Davis, R., and Jones, R. (2002). The effects of large-scale afforestation and climate change on water allocation in the Macquarie River catchment, NSW, Australia. *Journal of environmental management*, 65(4):369–381.

Hindmarsh, A. C. (1983). ODEPACK, a systematized collection of ODE solvers. In Stepleman, R. S., editor, *IMACS Transactions on Scientific Computation*, pages 55–64, Amsterdam, The Netherlands.

Hojati, M., Bector, C. R., and Smimou, K. (2005). A simple method for computation of fuzzy linear regression. *European Journal of Operational Research*, 166(1):172–184.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Hornberger, G. M. and Spear, R. C. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management*, 12:7–18.

Horsburgh, J. S., Tarboton, D. G., Hooper, R. P., and Zaslavsky, I. (2014). Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling & Software*, 52:62–73.

Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L. (2015). SPOTting model parameters using a ready-made Python package. *PloS one*, 10(12):e0145180.

Iman, R. L. and Conover, W. J. (2007). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*.

Jacquin, A. P. and Shamseldin, A. Y. (2007). Development of a possibilistic method for the evaluation of predictive uncertainty in rainfall-runoff modeling. *Water Resources Research*, 43(W04425):1–18.

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.

Jakeman, A. J., Littlewood, I. G., and Whitehead, P. G. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*, 117(1-4):275–300.

Jones, E., Oliphant, T., and Peterson, P. (2001). SciPy: Open source scientific tools for Python.

Kavetski, D. and Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research*, 46(W10511):1–27.

Kavetski, D. and Clark, M. P. (2011). Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. *Hydrological Processes*, 25:661–670.

Kavetski, D. and Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, 47(11):1–19.

Kavetski, D., Fenicia, F., and Clark, M. P. (2011). Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. *Water Resources Research*, 47(W05501):1–25.

Kavetski, D. and Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research*, 43(W03411):1–9.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, 42(3):1–9.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, 42(3):1–10.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, 42(W03S04):1–5.

Kokkonen, T. S. and Jakeman, A. J. (2001). A comparison of metric and conceptual approaches in rainfall-runoff modeling and its implications. *Water Resources Research*, 37(9):2345–2352.

Konikow, L. F. and Bredehoeft, J. D. (1992). Ground-water models cannot be validated. *Advances in Water Resources*, 15(1):75–83.

Kraft, P. (2012). *A hydrological programming language extension for integrated catchment models*. Phd thesis, University Giessen.

Kraft, P., Multsch, S., Vaché, K. B., Frede, H.-G., and Breuer, L. (2010). Using Python as a coupling platform for integrated catchment models. *Advances in Geosciences*, 27:51–56.

Kraft, P., Vaché, K. B., Frede, H.-G., and Breuer, L. (2011). CMF: A hydrological programming language extension for integrated catchment models. *Environmental Modelling & Software*, 26(6):828–830.

Kralisch, S., Krause, P., and David, O. (2005). Using the object modeling system for hydrological model development and application. *Advances In Geosciences*, 4:75–81.

Krause, P., Kralisch, S., and Flügel, W.-A. (2005). Model integration and development of modular modelling systems. *Advances In Geosciences*, 4:1–2.

Krueger, T., Freer, J. E., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., Butler, P., and Haygarth, P. M. (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, 46(W07516):1–17.

Kuczera, G., Kavetski, D., Franks, S. W., and Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, 331(1-2):161–177.

Kuczera, G., Renard, B., Thyer, M., and Kavetski, D. (2010). There are no hydrological monsters, just models and observations with large uncertainties! *Hydrological Sciences Journal*, 55(6):980–991.

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., and Hughes, A. (2013). Integrated environmental modeling: A vision and roadmap for the future. *Environmental Modelling & Software*, 39:3–23.

Le Phong, V. V., Kumar, P., Valocchi, A. J., and Dang, H.-V. (2015). GPU-based high-performance computing for integrated surfacesub-surface flow modeling. *Environmental Modelling & Software*, 73:1–13.

Leavesley, G. H., Markstrom, S. L., Restrepo, P. J., and Viger, R. J. (2002). A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling. *Hydrological processes*, 16:173–187.

Lee, H., McIntyre, N., Wheater, H. S., and Young, A. (2005). Selection of conceptual models for regionalisation of the rainfall-runoff relationship. *Journal of Hydrology*, 312(1-4):125–147.

Lee, H., McIntyre, N., Wheater, H. S., Young, A., and Wagener, T. (2004). Assessment of rainfall-runoff model structures for regionalisation purposes. In Webb, B., Arnell, N., Onof, C., MacIntyre, N., Gurney, R., and Kirby, C., editors, *Hydrology: science and practice for the 21st century, Volume 1. Proceedings of the British Hydrological Society International Conference*, pages 302–308, London. Imperial College.

Legates, D. R. and McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1):233–241.

Li, L., Xia, J., Xu, C.-Y., and Singh, V. P. (2010). Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models. *Journal of Hydrology*, 390(3-4):210–221.

Lilburne, L. R. R. and Tarantola, S. (2009). Sensitivity analysis of spatial models. *International Journal of Geographical Information Science*, 23(2):151–168.

Lin, Z. and Beck, M. B. (2007). On the identification of model structure in hydrological and environmental systems. *Water Resources Research*, 43(2):1–19.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201:272–288.

Liu, Y., Freer, J., Beven, K., and Matgen, P. (2009a). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, 367(1-2):93–103.

Liu, Y., Freer, J. E., Beven, K. J., and Matgen, P. (2009b). Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, 367(1-2):93–103.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H. (2012). A framework for benchmarking land models. *Biogeosciences*, 9(10):3857–3874.

MacKay, D. J. C. (2002). *Information theory, inference & learning algorithms*. Cambridge University Press, New York, NY, USA.

Madsen, H. (2000). Automatic calibration of a conceptual rainfallrunoff model using multiple objectives. *Journal of Hydrology*, 235(3-4):276–288.

Magee, B. (1973). *Popper*. William Collins Sons & Co, Glasgow.

Maier, H. R., Kapelan, Z., Kasprzyk, J., and Matott, L. S. (2015). Thematic issue on evolutionary algorithms in water resources. *Environmental Modelling & Software*, 69:222–225.

Mantovan, P. and Todini, E. (2006). Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology*, 330(1-2):368–381.

Mantovan, P., Todini, E., and Martina, M. (2007). Reply to comment by Keith Beven, Paul Smith and Jim Freer on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology". *Journal of Hydrology*, 338(3-4):319–324.

Mara, T. A., Tarantola, S., and Annoni, P. (2015). Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72:173–183.

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, 16(1):274.

Matott, L. S., Babendreier, J. E., and Purucker, S. T. (2009). Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resources Research*, 45(W06421):1–14.

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulations*.

Maxwell, T. and Costanza, R. (1997). An open geographic modeling environment. *Simulation*, 68(3):175–185.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, Austin, USA.

McMillan, H. K. and Clark, M. P. (2009). Rainfall runoff model calibration using informal likelihood measures within a Markov Chain Monte Carlo sampling scheme. *Water Resources Research*, 45(4):1–12.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., and Woods, R. A. (2011). Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure. *Hydrological Processes*, 25(4):511–522.

McMillan, H. K., Freer, J. E., Pappenberger, F., Krueger, T., and Clark, M. P. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, 24(10):1270–1284.

Meadows, D. H. (2009). *Thinking in systems - a primer*. Earthscan Ltd.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,*, 21(6):1087–1092.

Mishra, S. (2009). Uncertainty and sensitivity analysis techniques for hydrologic modeling. *Journal of Hydroinformatics*, 11(3-4):282–296.

Montanari, A. (2007). What do we mean by uncertainty ? The need for a consistent wording about uncertainty assessment in hydrology. *Hydrological Processes*, 21:841–845.

Montanari, A. and Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, 48(19):1–15.

Moore, R. J. (1985). The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal*, 30(2):273–297.

Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System Sciences*, 11(1):483–499.

Moore, R. V. and Tindall, C. I. (2005). An overview of the open modelling interface and environment (the OpenMI). *Environmental Science and Policy*, 8:279–286.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the Asabe*, 50(3):885–900.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Muleta, M. K. (2012). Improving model performance using season-based evaluation. *Journal of Hydrologic Engineering*, 17(1):191–200.

Nash, J. E. and Sutcliffe, J. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. Irwin, Chicago.

Newman, S. (2015). *Building microservices, designing fine-grained systems*. O'Reilly Media.

Nielsen, S. A. and Hansen, E. (1973). Numerical simulation of the rainfall-runoff process on a daily basis. *Nordic Hydrology*, 4(3):171 – 190.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer-Verlag New York, New York.

Nopens, I. (2010). *Modelleren en simuleren van biosystemen*. Course notes, Ghent, Belgium.

Nossent, J. (2012). *Sensitivity and uncertainty analysis in view of the parameter estimation of a SWAT model of the River Kleine Nete , Belgium*. Phd, Vrije Universiteit Brussel.

Nossent, J. and Bauwens, W. (2012a). Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol sensitivity analysis of a hydrological model. In *Geophysical Research Abstracts*, volume 14, Vienna, Austria. European Geosciences Union (EGU).

Nossent, J. and Bauwens, W. (2012b). Optimising the convergence of a Sobol' sensitivity analysis for an environmental model: Application of an appropriate estimate for the square of the expectation value and the total variance. In Seppelt, R., Voinov, A. A., Lange, S., and Bankamp, D., editors, *iEMSs 2012 - Managing Resources of a Limited Planet: Proceedings of the 6th Biennial Meeting of the International Environmental Modelling and Software Society*, pages 1080–1087, Leipzig, Germany. International Environmental Modelling and Software Society (iEMSs).

Nossent, J., Leta, O. T., and Bauwens, W. (2013). Assessing the convergence of a Morris-like screening method for a complex environmental model. In *7th International Conference on Sensitivity Analysis of Model Output*, Nice, France.

Nott, D. J., Marshall, L., and Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research*, 48(W12602):1–7.

Nuzzo, R. (2015). How scientists fool themselves and how they can stop. *Nature*, 526(7572):182–185.

Obled, C., Zin, I., and Hingray, B. (2009). Optimal space and time scales for parsimonious rainfall-runoff models. *La Houille Blanche*, 5:81–87.

Olivera, F., Valenzuela, M., Srinivasan, R., Choi, J., Cho, H., Koka, S., and Agrawal, A. (2006). ArcGIS-SWAT: A geodata model and GIS interface for SWAT. *Journal of the American Water Resources Association*, 42(2):295–309.

Omlin, M. and Reichert, P. (1999). A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling*, 115:45–59.

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, validation and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646.

Pappenberger, F. and Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research*, 42(5):1–8.

Pappenberger, F., Matgen, P., Beven, K. J., Henry, J., Pfister, L., and De Fraipont, P. (2006). Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources*, 29:1430–1449.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522:697–713.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). PyMC : Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4):1–81.

Peckham, S. D. (2008). Geomorphometry and spatial hydrologic modelling. In Hengl, T. and Reuter, H. I., editors, *Geomorphometry: Concepts, Software and Applications. Developments in Soil Science volume 33*, chapter 25, pages 377–393.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, N.Y.)*, 334(6060):1226–7.

Peters, N. E., Freer, J. E., and Beven, K. J. (2003). Modelling hydrologic responses in a small forested catchment (Panola Mountain, Georgia, USA): A comparison of the original and a new dynamic TOPMODEL. *Hydrological Processes*, 17(2):345–362.

Petersen, B. (2000). *Calibration, identifiability and optimal experimental design of activated sludge models*. PhD thesis, Ghent University.

Petersen, B., Gernaey, K., and Vanrolleghem, P. A. (2001). Practical identifiability of model parameters by combined respirometric-titrimetric measurements. *Water Science and Technology*, 43(7):347–355.

Petre, M. and Wilson, G. (2014). Code review for and by scientists. *arXiv preprint arXiv:1407.5648*, page 4.

Pfannerstill, M., Guse, B., and Fohrer, N. (2014). Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of Hydrology*, 510:447–458.

Pianosi, F., Sarrazin, F., and Wagener, T. (2015). A Matlab toolbox for global sensitivity analysis. *Environmental Modelling & Software*, 70:80–85.

Popper, K. (1959). *The logic of scientific discovery*. Hutchinson, London, England.

Prabhu, P., Jablin, T. B., Raman, A., Zhang, Y., Huang, J., Kim, H., Johnson, N. P., Liu, F., Ghosh, S., Beard, S., Oh, T., Zoufaly, M., Walker, D., and August, D. I. (2011). A survey of the practice of computational science. In *Proceedings of the 24th ACM/IEEE Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–12.

Pullar, D. (2004). SimuMap: A computational system for spatial modelling. *Environmental Modelling & Software*, 19:235–243.

R Core Development Team, R. (2008). R: A language and environment for statistical computing.

Refsgaard, J. C. (2004). Modelling guidelines - terminology and guiding principles. *Advances in Water Resources*, 27(1):71–82.

Refsgaard, J. C. and Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, 32(7):2189–2202.

Refsgaard, J. C. and Storm, B. (1995). MIKE SHE. In Singh, V. P., editor, *Computer Models of Watershed Hydrolog*, pages 809–846. Water Resources Publications, Colorado, USA.

Reggiani, P., Hassanizadeh, S. M., Sivapalan, M., and Gray, W. G. (1999). A unifying framework for watershed thermodynamics: constitutive relationships. *Advances in Water Resources*, 23(1):15–39.

Reggiani, P., Sivapalan, M., and Hassanizadeh, S. M. (1998). A unifying framework for watershed thermodynamics: Balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. *Advances in Water Resources*, 22(4):367–398.

Reichert, P. (1994). AQUASIM - A tool for simulation and data analysis of aquatic systems. *Water Science & Technology*, 30(2):21–30.

Reichert, P. (2003). *Umweltsystemanalyse: Identifikation von Modellen für Umweltsysteme und Schätzung der Unsicherheit von Modelprognosen*. Course notes, EAWAG, Zürich.

Reichert, P. and Mieleitner, J. (2009). Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research*, 45(10):1–19.

Reichert, P. and Omlin, M. (1997). On the usefulness of overparameterized ecological models. *Ecological Modelling*, 95(2-3):289–299.

Reichert, P. and Vanrolleghem, P. A. (2001). Identifiability and uncertainty analysis of the River Water Quality Model No. 1 (RWQM1). *Water Science & Technology*, 43(7):329–338.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, 47(W11516):1–21.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrological models. *Hydrology and Earth System Sciences*, 13(7):999–1018.

Reusser, D. E. and Zehe, E. (2011). Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity. *Water Resources Research*, 47(W07550):1–15.

Roche, D. G., Kruuk, L. E. B., Lanfear, R., and Binning, S. A. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, 13(11):e1002295.

Rodriguez-Fernandez, M., Mendes, P., and Banga, J. R. (2006). A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*, 83(2-3):248–265.

Romanowicz, R. J. and Beven, K. J. (2006). Comments on generalised likelihood uncertainty estimation. *Reliability Engineering & System Safety*, 91(10-11):1315–1321.

Romanowicz, R. J., Beven, K. J., and Tawn, J. A. (1994). Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. In Barnett, V. and Turkman, K. F., editors, *Statistics for the environment 2: Water Related Issues*, pages 297–317. John Wiley & Sons Ltd.

Rossum, G. (1995). Python reference manual. Technical report, Amsterdam, The Netherlands.

Rouhani, H., Willems, P., Wyseure, G., and Feyen, J. (2007). Parameter estimation in semi-distributed hydrological catchment modelling using a multi-criteria objective function. *Hydrological Processes*, 21:2998– 3008.

Rubarenzya, M. H., Willems, P., and Berlamont, J. (2007). Identification of uncertainty sources in distributed hydrological modelling: Case study of the Grote Nete catchment in Belgium. *Water SA*, 33(5):633–642.

Sadegh, M. and Vrugt, J. A. (2013). Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation. *Hydrology and Earth System Sciences*, 17(12):4831–4850.

Sadegh, M. and Vrugt, J. A. (2014). Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC). *Water Resources Research*, 50:6767–6787.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270.

Saltelli, A. and Bolado, R. (1998). An alternative way to compute Fourier amplitude sensitivity test (FAST). *Computational Statistics & Data Analysis*, 26:445–460.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis, The Primer*. John Wiley & Sons Ltd.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. Halsted Press, New York, NY, USA.

Salvadore, E., Bronders, J., and Batelaan, O. (2015). Hydrological modelling of urbanized catchments: a review and future directions. *Journal of Hydrology*, 529:62–81.

Schaefli, B. and Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, 2080(May):2075–2080.

Schlesinger, S., Crosbie, R. E., Gagné, R. E., Innis, G. S., Lalwani, C. S., and Loch, J. (1979). Terminology for model credibility. *Simulation*, 32(3):103–104.

Schmitz, O., Karssenberg, D., de Jong, K., de Kok, J.-L., and de Jong, S. M. (2013). Map algebra and model algebra for integrated model building. *Environmental Modelling & Software*, 48:113–128.

Schmitz, O., Salvadore, E., Poelmans, L., van der Kwast, J., and Karssenberg, D. (2014). A framework to resolve spatio-temporal misalignment in component-based modelling. *Journal of Hydroinformatics*, 16(4):850.

Schoups, G. and Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46(10):1–17.

Schoups, G., Vrugt, J. A., Fenicia, F., and van de Giesen, N. C. (2010). Corruption of accuracy and efficiency of Markov Chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models. *Water Resources Research*, 46(W10530):1–12.

Seppelt, R. and Richter, O. (2005). It was an artefact not the result: A note on systems dynamic model development tools. *Environmental Modelling & Software*, 20(12):1543–1548.

Shen, H. (2014). Interactive notebooks: Sharing the code. *Nature*, 515(7525):151–2.

Shou, W., Bergstrom, C. T., Chakraborty, A. K., and Skinner, F. K. (2015). Theory, models and biology. *eLife*, 4:1–4.

Shrestha, R. R. and Simonovic, P. S. (2010). Fuzzy nonlinear regression approach to stage-discharge analyses: Case study. *Journal of Hydrologic Engineering*, 15(1):49–56.

Sieber, A. and Uhlenbrook, S. (2005). Sensitivity analyses of a distributed catchment model to verify the model structure. *Journal of Hydrology*, 310(1-4):216–235.

Silberstein, R. P. (2006). Hydrological models are so good, do we still need data? *Environmental Modelling & Software*, 21(9):1340–1352.

Sin, G., Ödman, P., Petersen, N., Lantz, A. E., and Gernaey, K. V. (2008). Matrix notation for efficient development of first-principles models within PAT applications: Integrated modeling of antibiotic production with Streptomyces coelicolor. *Biotechnology and Bioengineering*, 101(1):153–171.

Sin, G., Villez, K., and Vanrolleghem, P. A. (2006). Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model. *Water Science & Technology*, 53(4-5):95–103.

Sivakumar, B. (2004). Dominant processes concept in hydrology: Moving forward. *Hydrological Processes*, 18(12):2349–2353.

Sivakumar, B. (2008). Dominant processes concept, model simplification and classification framework in catchment hydrology. *Stochastic Environmental Research and Risk Assessment*, 22(6):737–748.

Sivapalan, M., Blöschl, G., Zhang, L., and Vertessy, R. A. (2003). Downward approach to hydrological prediction. *Hydrological Processes*, 17(11):2101–2111.

Smith, T., Marshall, L., and Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, 528:29–37.

Sobol, I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 4(7):86–112.

Sobol, I. M. and Kucherenko, S. S. (2005). On global sensitivity analysis of quasi-Monte Carlo algorithms. *Monte Carlo Methods and Applications*, 11(1):83–92.

Soetaert, K. and Petzoldt, T. (2010a). Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. *Journal of Statistical Software*, 33(3):1–28.

Soetaert, K. and Petzoldt, T. (2010b). Solving differential equations in R: package deSolve. *Journal of Statistical Software*, 33(9):1–25.

Spanjers, H. and Vanrolleghem, P. A. (1995). Respirometry as a tool for rapid characterization of wastewater and activated sludge. *Water Science and Technology*, 31(2):105–114.

Spear, R. C. and Hornberger, G. M. (1980). Eutrophication in Peel inlet, II. Identification of critical uncertainties via Generalised Sensitivity Analysis. *Water Research*, pages 43–49.

Stedinger, J. R., Vogel, R. M., Lee, S. U., and Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, 44(W00B06):1–17.

SymPy Development Team (2014). SymPy: Python library for symbolic mathematics.

Tang, B. (1993). Orthogonal Array-Based Latin-Hypercube. *Journal of the American Statistical Association*, 88(424):1392–1397.

Tang, Y., Reed, P. M., van Werkhoven, K., and Wagener, T. (2007a). Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis. *Water Resources Research*, 43(6):1–14.

Tang, Y., Reed, P. M., Wagener, T., and van Werkhoven, K. (2007b). Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrology and Earth System Sciences*, 11(2):793–817.

Taylor, C. J., Pedregal, D. J., Young, P. C., and Tych, W. (2007). Environmental time series analysis and forecasting with the Captain toolbox. *Environmental Modelling & Software*, 22:797–814.

Ternbach, B. M. A. (2005). *Modeling based process development of fed-batch bioprocesses: L-Valine production by corynebacterium glutamicum.* PhD thesis, Aachen University.

Thiemann, M., Trosset, M., Gupta, H. V., and Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, 37(10):2521–2535.

Todini, E. (2007). Hydrological catchment modelling: past, present and future. *Hydrology and Earth System Sciences*, 11(1):468–482.

Tomassini, L., Reichert, P., Künsch, H. R., Buser, C., Knutti, R., and Borsuk, M. E. (2009). A smoothing algorithm for estimating stochastic, continuous time model parameters and its application to a simple climate model. *Applied Statistics*, 58(5):679–704.

Tripp, D. R. and Niemann, J. D. (2008). Evaluating the parameter identifiability and structural validity of a probability-distributed model for soil moisture. *Journal of Hydrology*, 353(1-2):93–108.

Uhlenbrook, S., Seibert, J., Leibundgut, C., and Rodhe, A. (1999). Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure. *Hydrological Sciences*, 44(5):779–797.

Usher, W., Herman, J., Hadka, D., Xantares, X., Bernardoct, B., and Mutel, C. (2015). SALib: Sensitivity Analysis Library in Python (Numpy). Contains Sobol, Morris, Fractional factorial and FAST methods.

Vaché, K. B. and McDonnell, J. J. (2006). A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resources Research*, 42(2):W02409.

Van Daele, T., Van Hauwermeiren, D., Ringborg, R., Heintz, S., Van Hoey, S., Gernaey, K. V., and Nopens, I. (2015a). A case study on robust optimal experimental design for model calibration of omega transaminase. In *Fifth European Process Intensification Conference*, Nice, France.

Van Daele, T., Van Hoey, S., Gernaey, K. V., Krühne, U., and Nopens, I. (2015b). A numerical procedure for model identifiability analysis applied to enzyme kinetics. In Gernaey, K. V., Huusom, J. K., and Gani, R., editors, *12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering*.

Van Daele, T., Van Hoey, S., and Nopens, I. (2015c). pyIDEAS: an open source python package for model analysis. In Gernaey, K. V., Huusom, J. K., and Gani, R., editors, *12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering*, Copenhagen, Denmark.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13:22–30.

Van Eerdenbrugh, K., Van Hoey, S., and Verhoest, N. (2016a). Identification of consistency in rating curve data: Bidirectional Reach (BReach). *Water Resources Research*, (in preparation).

Van Eerdenbrugh, K., Verhoest, N. E. C., and Van Hoey, S. (2016b). BReach (Bidirectional Reach): A methodology for data consistency assessment applied on a variety of rating curve data. In *River Flow 2016 - Eight international conference on fluvial hydraulics*.

van Griensven, A. and Bauwens, W. (2003). Multiobjective autocalibration for semidistributed water quality models. *Water Resources Research*, 39(12):1–9.

van Griensven, A. and Meixner, T. (2007). A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models. *Journal of Hydroinformatics*, 9(4):277–291.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., and Srinivasan, R. (2006). A global sensitivity analysis tool for the parameters of multi-variable catchment models. *Journal of Hydrology*, 324(1-4):10–23.

Van Hoey, S. (2008). *Invloed van ruimtelijke neerslaginterpollatietechnieken op de hydrologische modellering van het Nzoia stroomgebied*. Msc thesis, University Ghent.

Van Hoey, S., Balemans, S., Nopens, I., and Seuntjens, P. (2015a). Hydropy: Python package for hydrological time series handling based on Python Pandas. In *European Geoscience Union General Assembly (PICO session on open source in hydrology)*, Vienna, Austria.

Van Hoey, S., Dams, J., Seuntjens, P., and Nopens, I. (2014a). Next-generation tools m.b.t. hydrometrie, hydrologie en hydraulica in het operationeel waterbeheer, Fase 1: analyse, Perceel 1: De Maarkebeek, description standard evaluation method for hydrological models. Technical report, Vlaamse Milieu Maatschappij.

Van Hoey, S., Nopens, I., van der Kwast, J., and Seuntjens, P. (2015b). Dynamic identifiability analysis-based model structure evaluation considering rating curve uncertainty. *Journal of Hydrologic Engineering*, 20(5):1–17.

Van Hoey, S., Seuntjens, P., van der Kwast, J., de Kok, J.-L., Engelen, G., and Nopens, I. (2011). Flexible framework for diagnosing alternative model structures through sensitivity and uncertainty analysis. In Chan, F., Marinova, D., and Anderssen, R. S., editors, *MODSIM2011, 19th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand*, pages 3924–3930. Modelling and Simulation Society of Australia and New Zealand (MSSANZ).

Van Hoey, S., Seuntjens, P., van der Kwast, J., and Nopens, I. (2014b). A qualitative model structure sensitivity analysis method to support model selection. *Journal of Hydrology*, 519:3426–3435.

Van Hoey, S., Vansteenkiste, T., Pereira, F., Nopens, I., Seuntjens, P., Willems, P., and Mostaert, F. (2012). Effect of climate change on the hydrological regime of navigable water courses in Belgium: Subreport 4 Flexible model structures and ensemble evaluation. Technical report, Waterbouwkundig Laboratorium, Antwerpen, België.

Vandenberghe, S. (2012). *Copula-based models for generating design rainfall*. PhD thesis, Ghent University.

VanderPlas, J. (2014). Frequentism and Bayesianism: A Python-driven Primer. In *proceedings of the 13th Python in science conference (Scipy 2014)*, pages 1–9.

Vanhooren, H., Meirlaen, J., Amerlinck, Y., Claeys, F., Vangheluwe, H., and Vanrolleghem, P. A. (2003). Modeling biological wastewater treatment. *Journal of Hydroinformatics*, 5:27–50.

Vanrolleghem, P. A. and Keesman, K. J. (1996). Idenfication of biodegration models under model and data uncertainty.

Vanrolleghem, P. A., Mannina, G., Cosenza, A., and Neumann, M. B. (2015). Global sensitivity analysis for urban water quality modelling: Terminology, convergence and comparison of different methods. *Journal of Hydrology*, 522:339–352.

Vanrolleghem, P. A., Sin, G., and Gernaey, K. V. (2004). Transient response of aerobic and anoxic activated sludge activities to sudden substrate concentration changes. *Biotechnology and bioengineering*, 86(3):277–90.

Vanrolleghem, P. A., Spanjers, H., Petersen, B., Ginestet, P., and Takacs, I. (1999). Estimating (combinations of) Activated Sludge Model No. 1 parameters and components by respirometry. In *Water Science and Technology*, volume 39, pages 195–214.

Vanrolleghem, P. A., Van Daele, M., and Dochain, D. (1995). Practical identifiability of a biokinetic model of activated sludge respiration. *Water Research*, 29(11):2561–2570.

Vansteenkiste, T., Pereira, F., Willems, P., and Mostaert, F. (2011). Effect of climate change on the hydrological regime of navigable water courses in Belgium: Subreport 2 - Climate change impact analysis by conceptual models. Versie 1_0. , 706_18. Technical report, Waterbouwkundig Laboratorium & K.U.Leuven, Antwerpen, België.

Verborgh, R. and De Wilde, M. (2013). *Using OpenRefine*. Packt Publishing, 1st edition.

Verwaeren, J., der Weeën, P., and De Baets, B. (2015). A search grid for parameter optimization as a byproduct of model sensitivity analysis. *Applied Mathematics and Computation*, 261:8–38.

Villa, F. (2001). Integrating modelling architecture: A declarative framework for multi-paradigm, multi-scale ecological modelling. *Ecological Modelling*, 137(1):23–42.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. A., and Buytaert, W. (2015). Web technologies for environmental Big Data. *Environmental Modelling & Software*, 63:185–198.

Voinov, A., Fitz, C., Boumans, R., and Costanza, R. (2004). Modular ecosystem modeling. *Environmental Modelling & Software*, 19:285–304.

Voinov, A. and Shugart, H. H. (2013). "Integronsters", integral and integrated modeling. *Environmental Modelling & Software*, 39:149–158.

Vrugt, J. A. (2015). Markov Chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, accepted.

Vrugt, J. A. and Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, 43:1–15.

Vrugt, J. A. and Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, 49(7):4335–4345.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A. (2008a). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov Chain Monte Carlo simulation. *Water Resources Research*, 44(W00B09):1–15.

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A. (2008b). Response to comment by Keith Beven on "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling". *Stochastic Environmental Research and Risk Assessment*, 23(7):9–10.

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S. (2001a). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, 5(1):13–26.

Wagener, T. and Kollat, J. B. (2007). Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. *Environmental Modelling & Software*, 22(7):1021–1033.

Wagener, T., Lees, M. J., and Wheater, H. S. (2001b). A toolkit for the development and application of parsimonious hydrological models. In Singh, V. P., Frevert, D., and Meyer, D., editors, *Mathematical models of small watershed hydrology - Volume 2*, pages 1–34. Water Resources Publications.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, 17(2):455–476.

Wagener, T., van Werkhoven, K., Reed, P. M., and Tang, Y. (2009). Multiobjective sensitivity analysis to understand the information content in streamflow observations for distributed watershed modeling. *Water Resources Research*, 45(W02501):1–5.

Wagener, T., Wheater, H., and Gupta, H. V. (2004). *Rainfall-runoff Modelling in gauged and ungauged catchments*. Imperial College Press, London, England.

Wagener, T. and Wheater, H. S. (2002). A generic framework for the identification of parsimonious rainfall-runoff models. In Rizzoli, A. E. and Jakeman, A. J., editors, *iEMSs 2002 International Congress: "Integrated Assessment and Decision Support". Proceedings of the 1st biennial meeting of the International Environmental Modelling and Software Societey*, pages 434–439, Lugano, Switzerland.

Welsh, W. D., Vaze, J., Dutta, D., Rassam, D., Rahman, J. M., Jolly, I. D., Wallbrink, P., Podger, G. M., Bethune, M., Hardy, M. J., Teng, J., and Lerat, J. (2013). An integrated modelling framework for regulated river systems. *Environmental Modelling & Software*, 39:1–22.

Wesseling, C. G., Karssenberg, D., Van Deursen, W. P. A., and Burrough, P. A. (1996). Integrating dynamic environmental models in GIS: The development of a dynamic modelling language. *Transactions in GIS*, 1:40–48.

Westerberg, I. K., Guerrero, J., Seibert, J., Beven, K. J., and Halldin, S. (2011a). Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 613(25):603– 613.

Westerberg, I. K., Guerrero, J., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y. (2011b). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7):2205–2227.

Willems, P. (2009). A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. *Environmental Modelling & Software*, 24(3):311–321.

Willems, P. (2014). Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes - Part 1: Step-wise model-structure identification and calibration approach. *Journal of Hydrology*, 510:578–590.

Willems, P., Mora, D., Vansteenkiste, T., Taye, M. T., and Van Steenbergen, N. (2014). Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes - Part 2: Intercomparison of models and calibration approaches. *Journal of Hydrology*, 510:591–609.

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., and Wilson, P. (2014). Best practices for scientific computing. *PLoS biology*, 12(1):1–7.

Winsemius, H. C., Schaefli, B., Montanari, a., and Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(W12422):1–15.

Wolfs, V., Meert, P., and Willems, P. (2015). Modular conceptual modelling approach and software for river hydraulic simulations. *Environmental Modelling & Software*, 71:60–77.

Wriedt, G. and Rode, M. (2006). Investigation of parameter uncertainty and identifiability of the hydrological model WaSiM-ETH. *Advances in Geosciences*, 9:145–150.

Xu, D.-M., Wang, W.-C., Chau, K.-W., Cheng, C.-T., and Chen, S.-Y. (2013). Comparison of three global optimization algorithms for calibration of the Xinanjiang model parameters. *Journal of Hydroinformatics*, 15(1):174–193.

Yang, J. (2011). Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. *Environmental Modelling & Software*, 26(4):444–457.

Young, P. C. (2003). Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. *Hydrological Processes*, 17(11):2195–2217.

Young, P. C., McKenna, P., and Bruun, J. (2001). Identification of non-linear stochastic systems by state dependent parameter estimation. *International Journal of control*, 74(18):1837–1857.

Curriculum Vitae

✉ stijnvanhoey@gmail.com @SVanHoey github.com/stijnvanhoey

education

study program

since 2008 **Ph.D.** candidate in Bioscience Engineering — Biomath, Ghent University & VITO
Toolbox for model structure evaluation and selection
Development of methodologies for improved model structure selection in environmental applications with the focus on water management
supervised by Prof dr ir Piet Seuntjens and Prof dr ir Ingmar Nopens

2007 **Erasmus exchange programme** — BOKU, Vienna
M.Sc program *Kulturtechnik und Wasserwirtschaft*

2003-2008 **B.Sc/M.Sc. Bioscience Engineering in Land Management and Forestry** — Ghent University
option water- and soil management

1997-2003 **High school** — Heilig-Maagdcollege Dendermonde
Science and Mathematics

additional courses

2014 **Regression Analysis and Advanced Regression Topics** — FLAMES summer school, Belgium

2011 **Getting started with high-performance computing** — Doctoral Schools UGent, Belgium

2010 **First Annual Catchment Science Summer School** — Aberdeen, Scotland

2010 **Inverse Modelling in Earth and Environmental Sciences** — summer school KULeuven, Belgium

2010 **Effective Scientific Communication** — Doctoral Schools UGent, Belgium

2008 **Map Algebra and Dynamic Modelling with PCRaster** — VITO, Belgium

2007 **International Workshop on Soil and Water Assessment Tool (SWAT)** — UNESCO-IHE, Netherlands

2007 **German basic course** — University Language Centre Ghent, Belgium

employment

2013-current **Teaching assistant** — Biomath, Ghent University

2012-2013 **Scientific researcher** — Biomath, Ghent University

2008-2012 **Phd candidate** — Biomath, Ghent University & VITO

teaching

courses

2015 **Scientific computing (Matlab)** Ghent University
assisting during exercise lectures, B.Sc. Bioscience Engineering

2015 **Scientific computing with Python** Antea Group
2-day intensive course in collaboration with Joris Van den Bossche

2015 **Beslissingsondersteunende technieken** Ghent University
development course material and teaching with IPython notebook, Manama environmental sanitation

2014, 2015 **Modelling and simulation of biosystems** Ghent University
exercise lectures, B.Sc. Bioscience Engineering

2014, 2015 **Modelling and control of waste water treatment systems** Ghent University
exercise lectures, M.Sc. Bioscience Engineering

2014, 2015 **Integrated modelling and design of basin management plans** Ghent University
exercise lectures, Technology for Integrated Water Management

2013 **Advanced statistical methods, module 1: The package R.** IVPV, Ghent University
assisting during exercise lectures

2013 **Introduction to uncertainty assessment of environmental models** UNESCO-IHE, Delft
guest lecture, Environmental and Global Change: Uncertainty and Risk Assessment

2012, 2013, 2014, 2015 **Introduction to hydrological modelling** Ghent University
guest lecture, Technology for Integrated Water Management

2010, 2011, 2014, 2015 **Process control** Ghent University
assisting during exercise lectures, M.Sc. Bioscience Engineering

2009 **Process control** Ghent University
exercise lectures, M.Sc. Bioscience Engineering

tutorship

2014-2015 **Model-based scenario analysis for optimizing WWTPs operation in Bogota, Colombia**
M.Sc. Technology for Integrated Water Management
Elke Smets

2014-2015 **Bepaling van de meest representatieve meetplaats van het grondwaterpeil bij infrastructuurwerken** B.Sc. project Bioscience Engineering
Marjolein Minne, Eline Mauroo, Jasmine Heyse en Karen Roels

2013-2014 **Model-based analysis of electrochemical systems for sulfide recovery**
M.Sc. Bioscience Engineering
Robin Michelet
Master Thesis award: Environmental Prize Arcelor-Mittal with Indaver

2013-2014 **Linking the carbon biokinetics of activated sludge to the operational wastewater treatment conditions** M.Sc. Bioscience Engineering
Stijn Decubber

conference committee

2013 **3rd Young Water Professionals BENELUX conference, Luxembourg** Organizing committee

2013 **2nd OpenWater symposium and workshops, Belgium** Scientific committee

publications

articles in peer-reviewed journal

Identification of consistency in rating curve data: Bidirectional Reach (BReach)
K. Van Eerdenbrugh, **S. Van Hoey**, N. Verhoest
Water Resources Research*, in revision, 2016*

Anonymous peer assessment in higher education: exploring its effect on interpersonal variables and students' preferred type of teacher assessment and feedback
T. Rotsaert, T. Schellens, B. De Wever, A. Raes, **S. Van Hoey**
PLOS ONE*, submitted, 2016*

Validation of a microalgal growth model accounting with inorganic carbon and nutrient kinetics for wastewater treatment
B. Decostere, J. De Craene, **S. Van Hoey**, H. Vervaeren, S. Nopens
Chemical Engineering Journal*, 285, pp. 189–197, 2016*

Sensitivity of water stress in a two-layered sandy grassland soil to variations in groundwater depth and soil hydraulic parameters
M. Rezaei, P. Seuntjens, I. Joris, W. Boënne, **S. Van Hoey**, P. Campling, W. Cornelis
Hydrology and Earth System Sciences*, 20, pp. 487–503, 2016*

Dynamic identifiability analysis-based model structure evaluation considering rating curve uncertainty
S. Van Hoey, I. Nopens, J. van der Kwast, P. Seuntjens
Journal of Hydrologic Engineering*, 20.5, pp. 1–17, 2015*

From the affinity constant to the half-saturation index: understanding conventional modeling concepts in novel wastewater treatment processes
M. Arnaldos Orts, Y. Amerlinck, U. Rehman, T. Maere, **S. Van Hoey**, W. Naessens, I. Nopens
Water Research*, 70, pp. 458–470, 2015*

Sensitivity analysis of a soil hydrological model for estimating soil water content in a two-layered sandy soil for irrigation management purposes
M. Rezaei, P. Seuntjens, I. Joris, W. Boënne, **S. Van Hoey**, W. Cornelis
Vadose Zone Journal*, accepted, 2014*

A qualitative model structure sensitivity analysis method to support model selection
S. Van Hoey, P. Seuntjens, J. Kwast, I. Nopens
Journal of Hydrology*, 519, pp. 3426–3435, 2014*

A GLUE uncertainty analysis of a drying model of pharmaceutical granules
S. Mortier, **S. Van Hoey**, K. Cierkens, K. V. Gernaey, P. Seuntjens, B. De Baets, T. De Beer, I. Nopens
European Journal of Pharmaceutics and Biopharmaceutics*, 85.3, pp. 984–995, 2013*

peer-reviewed conferences/proceedings

BReach (Bidirectional Reach): a methodology for data consistency assessment applied on a variety of rating curve data
K. Van Eerdenbrugh, N. E. C. Verhoest, **S. Van Hoey**
River Flow 2016, Eight international conference on fluvial hydraulics. Saint Louis, Missouri, USA, 2016

A numerical procedure for model identifiability analysis applied to enzyme kinetics
T. Van Daele, **S. Van Hoey**, K. V. Gernaey, U. Krühne, I. Nopens
12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering ed. by K. V. Gernaey, J. K. Huusom, and R. Gani. Copenhagen, Denmark, June 2015

pyIDEAS: an open source python package for model analysis
T. Van Daele, **S. Van Hoey**, I. Nopens
12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering ed. by K. V. Gernaey, J. K. Huusom, and R. Gani. Copenhagen, Denmark, June 2015

Optimizing Hydrus 1D for irrigation management purposes in sandy grassland
M. Rezaei, I. Joris, W. Boënne, **S. Van Hoey**, P. Seuntjens, W. Cornelis
Water technology and management : proceedings of the 2nd European symposium ed. by L. Bastiaens. Leuven, Belgium, Nov. 2013

Influence of uncertainty analysis methods and subjective choices on prediction uncertainty for a respriometric case
K. Cierkens, **S. Van Hoey**, B. De Baets, P. Seuntjens, I. Nopens
International congress on environmental modelling and software : managing resources of a limited planet ed. by R. Seppelt, A. A. Voinov, S. Lange, and D. Bankamp. Leipzig, Germany, July 2012

Classification of quality attributes for software systems in the domain of integrated environmental modelling
M. Naeem, **S. Van Hoey**, P. Seuntjens, W. Boënne, P. Viaene
International congress on environmental modelling and software : managing resources of a limited planet ed. by R. Seppelt, A. A. Voinov, S. Lange, and D. Bankamp. Leipzig, Germany, July 2012

Comparison of uncertainty analysis methods for bioprocess modelling
K. Cierkens, **S. Van Hoey**, B. De Baets, P. Seuntjens, I. Nopens
Communications in agricultural and applied biological sciences. Leuven, Belgium, Feb. 2012

Flexible framework for diagnosing alternative model structures through sensitivity and uncertainty analysis
S. Van Hoey, P. Seuntjens, J. Van der Kwast, J. De Kok, G. Engelen, I. Nopens
MODSIM 2011 : 19th international congress on modelling and simulation ed. by F. Chan, D. Marinova, and R. S. Anderssen. Perth, WA, Australia, Dec. 2011

conference presentations

A phreatic groundwater level indicator: from monthly status assessment to daily forecasts
G. Heuvelmans, A. Fronhoffs, **S. Van Hoey**, K. Foncke, R. De Sutter, I. Rocabado, L. Candela, D. D'hont
42nd IAH International Congress, AQUA2015. Rome, Italy, 2015

A case study on robust optimal experimental design for model calibration of omega transaminase
T. Van Daele, D. Van Hauwermeiren, R. Ringborg, S. Heintz, **S. Van Hoey**, K. V. Gernaey, I. Nopens
Fifth European Process Intensification Conference. Nice, France, 2015

Hydropy: Python package for hydrological time series handling based on Python Pandas
S. Van Hoey, S. Balemans, I. Nopens, P. Seuntjens
European Geoscience Union General Assembly (PICO session on open source in hydrology). Vienna, Austria, 2015

Interactive model evaluation tool based on IPython notebook
S. Balemans, **S. Van Hoey**, I. Nopens, P. Seuntjens
European Geoscience Union General Assembly (PICO session on open source in hydrology). Vienna, Austria, 2015

Application of model uncertainty analysis on the modelling of the drying behaviour of single pharmaceutical granules
S. Mortier, **S. Van Hoey**, K. Cierkens, T. De Beer, K. V. Gernaey, I. Nopens
Zestiende Forum der Farmaceutische wetenschappen. Blankenberge, Belgium, 2012

Model structure identification based on ensemble model evaluation
S. Van Hoey, J. Kwast, I. Nopens, P. Seuntjens, F. Pereira
European Geoscience Union General Assembly. Vienna, Austria, 2012

Model calibration as a combined process of parameterization and model structure identification
S. Van Hoey, I. Nopens, P. Seuntjens
Mini-symposium on Model Structure. Delft, The Netherlands, 2011

Selecting the optimal spatial detail and process complexity for modeling environmental systems
S. Van Hoey, I. Nopens, G. Engelen, J. Kwast, J. Kok, P. Seuntjens
Looking at Catchments in Colors, EGU Leonardo. Luxemburg, GD Luxemburg, 2010

Selecting the appropriate spatial detail and process complexity of the hydrological representation when modeling environmental systems
S. Van Hoey, P. Seuntjens, I. Nopens, J. Kwast, J. Kok
Hydrology Conference. San Diego, USA, 2010

Modeling micropollutant fate at the catchment scale: from science to practice
P. Seuntjens, N. Desmet, K. Holvoet, A. Van Griensven, **S. Van Hoey**, X. Tang, I. Nopens
European Geoscience Union General Assembly. Vienna, Austria, 2009

conference posters

Calibration and analysis of a direct contact membrane distillation (DCMD) model using the GLUE method
I. P. Hitsov, **S. Van Hoey**, L. Eykens, T. Maere, K. De Sitter, C. Dotremont, I. Nopens
IWA World Water congress, 9th, Abstracts. Lisbon, Portugal, 2014

Python package for model STructure ANalysis (pySTAN)
S. Van Hoey, J. Kwast, I. Nopens, P. Seuntjens
OpenWater, 2nd Symposium and workshops, Abstracts. Etterbeek, Belgium, 2013

Python package for model STructure ANalysis (pySTAN)
S. Van Hoey, J. Kwast, I. Nopens, P. Seuntjens
European Geoscience Union General Assembly. Vienna, Austria, 2013

Application of model uncertainty analysis on the modelling of the drying behaviour of single pharmaceutical granules
S. Mortier, **S. Van Hoey**, K. Cierkens, T. De Beer, K. V. Gernaey, I. Nopens
Pan-European QbD & PAT Science Conference, 5th, Abstracts. Ghent, Belgium, 2012

GIS based systems for agrochemicals and nutrients in river catchments: Tools to support decision making from regional to large scale
Seuntjens P. Broekx S. Kwast J. Haest P. J. **S. Van Hoey**
Studiedag gewasbeschermingsmiddelen. Brussels, Belgium, 2010

scientific consulting

Uitbreiding van de grondwaterstandsindicator tot een voorspeller van freatische grondwaterstanden
K. Foncke, **S. Van Hoey**, K. Wildemeersch, I. Nopens
Vlaamse Milieu Maatschappij, Brussels, Belgium, 2015

Model Onzekerheden: Methodologische Plan van Aanpak
M. Heredia, **S. Van Hoey**, I. Rocabado, I. Nopens
Waterbouwkundig Laboratorium, Antwerpen, Belgium, 2014

Next-generation tools m.b.t. hydrometrie, hydrologie en hydraulica in het operationeel waterbeheer Fase 1: analyse, Perceel 1: De Maarkebeek, Standard Evaluation Hydrological Models: G2G application
J. Dams, **S. Van Hoey**, P. Seuntjens, I. Nopens
Vlaamse Milieu Maatschappij, Brussels, Belgium, 2014

Next-generation tools m.b.t. hydrometrie, hydrologie en hydraulica in het operationeel waterbeheer Fase 1: analyse, Perceel 1: De Maarkebeek, Description Standard Evaluation Method Hydrological Models
S. Van Hoey, J. Dams, P. Seuntjens, I. Nopens
Vlaamse Milieu Maatschappij, Brussels, Belgium, 2014

Effect of climate change on the hydrological regime of navigable water courses in Belgium: Subreport 4, Flexible model structures and ensemble evaluation. WL Rapporten, 706 – 18
S. Van Hoey, T. Vansteenkiste, F. Pereira, I. Nopens, P. Seuntjens, P. Willems, F. Mostaert
Waterbouwkundig Laboratorium, Antwerp, Belgium, 2012