THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Cotutelle internationale avec :
Ghent University (UGent), Belgium

Présentée et soutenue par :
Ratha SOR
Le Lundi 10 juillet 2017

Titre :
Modélisation des changements spatio-temporels des communautés de macroinvertébrés benthiques dans les rivières d'Asie et d'Europe

---

ED SEVAB : Écologie, biodiversité et évolution

Unité de recherche :
Évolution et Diversité Biologique (UT3 Paul Sabatier, France)
Aquatic Ecology (UGent, Belgium)

Directeur(s) de Thèse :
Prof Sovan LEK (Université Toulouse 3 Paul Sabatier, France)
Prof Peter GOETHALS (Ghent University, Belgium)
Dr Pieter BOETS (Ghent University, Belgium)

Rapporteurs :
Prof Philippe USSEGLIO-POLATERA (University of Lorraine, France)
Dr Nam SO (Mekong River Commission, Laos PDR)
Prof. Olivier THAS (Ghent University, Belgium)

Autre(s) membre(s) du jury :
Prof Michele TACKX (Université Toulouse 3 Paul Sabatier, France)
Prof Young-Seuk PARK (Kyung Hee University, Republic of Korea)
Prof Sovan LEK (Université Toulouse 3 Paul Sabatier, France)
Dr Saveng ITH (Royal University of Phnom Penh, Cambodia)
Prof Peter GOETHALS (Ghent University, Belgium)
Prof Olivier THAS (Ghent University, Belgium)
Dr Pieter BOETS (Ghent University, Belgium)

Modelling spatio-temporal changes of benthic macroinvertebrate communities in Asian and European rivers

MSc Ratha SOR

Supervisors:

Prof Dr Sovan LEK

Laboratoire Evolution & Diversité Biologique, UMR 5174, Université Paul Sabatier – Toulouse III, 118 route de Narbonne, 31062 Toulouse cédex 4 – France

Prof Dr Peter GOETHALS

Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure links 653, B9000 Ghent, Belgium

Co-supervisor:

Dr Pieter BOETS

Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000 Ghent, Belgium

Thesis submitted in fulfilment of the requirements for the degree of Doctor (PhD) in Ecology, Diversity and Biology (Université Paul Sabatier – Toulouse III) and in Applied Biological Sciences (Ghent University).

This research was performed at:

Laboratoire Evolution & Diversité Biologique (2 years), UMR 5174, Université Paul Sabatier – Toulouse III, 118 route de Narbonne, 31062 Toulouse cédex 4 – France.

Laboratory for Environmental Toxicology and Aquatic Ecology (1 year), Department Applied Ecology and Environmental Biology, Faculty of Bioscience Engineering, Ghent University, Campus Coupure building F, Coupure links 653, B9000 Ghent, Belgium.

Please refer to this work as follows:

Sor R., 2017. Modelling spatio-temporal changes of benthic macroinvertebrate communities in Asian and European rivers. Joint-PhD thesis, University of Toulouse/Ghent University, Toulouse, France.

Acknowledgements

Firstly, I would like to express my sincere thanks and profound gratitude to my thesis directors Prof Dr Sovan LEK, Prof Dr Peter GOETHALS and Dr Pieter BOETS for their scientific support. They have a huge pile of practical, ecological and theoretical knowledge. Their guidance, advice, suggestions and encouragements allowed me to survive during this scientific journey. I thank you all very much for your precious time spent on reading and correcting my work and discussing with me with an open mind. I also highly appreciate the freedom I received and the possibility I got to begin my PhD research project.

I am grateful to my thesis referees, jury members and all reading committees (Prof Philippe USSEGLIO-POLTERA, Prof Michele TACKX, Prof Young-Seuk PARK, Dr Nam SO, Prof Olivier THAS and Dr Saveng ITH) for their acceptance to review and evaluate this thesis.

My grateful appreciation is due to lecturers and senior research staff from the Faculty of Science, Royal University of Phnom Penh (RUPP), who always supported and encouraged me to continue my PhD. They include Prof MEAK Kamerane, Lecturers: THAO Sokunthia, UK On Norong, THAI Iengly, HAP Sophorn, ITH Saveng, PHAUK Sophany, MEAS Senghun and KHEAM Sokha, Nicholas SOUTER and other lecturers in the Department of Biology (RUPP).

Many thanks go to technical, administrative and management staff in EDB lab (UPS-TLS3): Nicole Hommet, Linda Jalabert, Dominique Pantalacci, and in the Environmental Toxicology and Aquatic Ecology lab: Sigrid Schuermans, Marianne Van den Hove, Sylvie Bauwens, Veerle Lambert, Nancy De Saeyer. They facilitated all of administrative procedures during my stay in each laboratory.

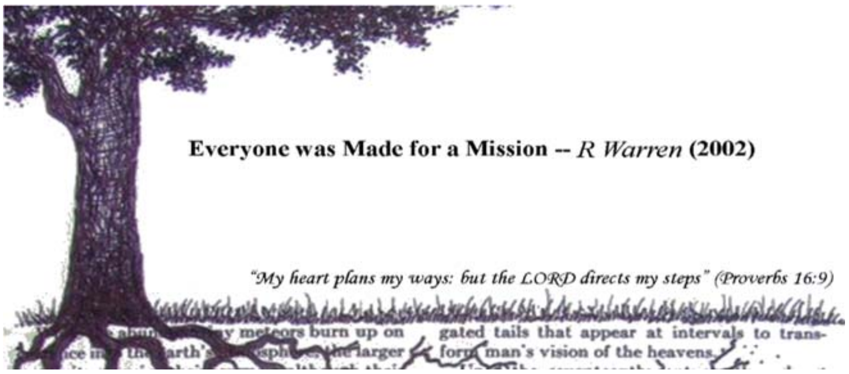
Warm thanks also go my AQUAECO team leaders, colleagues and friends: Gaël Grenouillet, Sébastien Brosse, Loïc Tudesque, Mathieu Chevalier, Pengbun Ngor, Ratha Chea, Ratha Seng, Kong Heng, Bunyeth Chan, Phanith Long, Sophea Iv, Sophany Soeun, Chuanbo Guo, Chao Wang, Mingli Lin, Kevin Cilleros, Marine Cambon, Jade Bruxaux, Lucie Kuczynski and other Cambodian friends. You all kept me a live with fun and food! Loïc, thank you very much for your helps with GIS, your editing my French thesis summary and your frequent friendly conversations. Sophea, highly appreciated for your help translating my thesis summary into French. Thank you also my friends from UGent and from the Environmental Toxicology and Aquatic Ecology lab: Niels De Troyer, Wout Van Echelpoel, Sacha Gobeyn, Jana Van Butsel, Minar Naomi Damanik Ambarita, Marie Anne Eurie Forio, Tuan Long Ho, Selamawit Negassa Chawaka, Natalia Carolina Donoso Pantoja, Rubén Jerves Cobo, Daniel Mercado Garcia, Thi Hanh Tien Nguyen, Tri Trương Trịnh Từ, Stijn Bruneel, Nathalie Claire Paracueles, Channy Chim, Saosometh Chhith and other friendly people. Our lunch was always relaxing and funny!

Financial means are very crucial to make this scientific mission possible. I am therefore grateful to the Erasmus Mundus LOTUS Unlimited project (European Commission) and the Belmont forum TLSCC projects for providing the chance to do this research. And also, I would like to thank the Mekong River Commission (MRC) and the Flemish Environment Agency (VMM) for the opportunity to study their samples.

My thankful appreciation owe to Prof Sovan LEK and Mrs Sithan LEK for their hospitality, kindness, care, and delicious food. They consider me like their own child, and their house feels like a home for me in Toulouse. I do enjoy gardening (especially during the weekend!), shopping and travelling with them. Many thanks to my second parents and their family in Belgium and my second mum (Judy Santmire) in the USA for their warm welcome, treats and talks during my stressing moments, and to brothers and sisters in churches (in Cambodia, France, and Belgium) for their supportive prayers. Highest thanks be to God for he has given me this wonderful scientific mission and for his wisdom and blessing so that I can complete this assignment with perseverance and joy.

Grateful thanks is extended to my beloved parents, brother and sister. It is blessing to be born and grow up in such a warm and lovely family.

កូនសូមអរគុណលោកពុក អ្នកម្ដាយជាអនេក សម្រាប់ការចិញ្ចឹមបីបាច់ថែរក្សា និងទំនុកបម្រុងគ្រប់បែបយ៉ាង។ អរគុណបងប្រុស និងបងស្រី ដែលមើលថែ និងការពារប្អូនតាំងពីតូចជាមួយពុកម៉ែ។ សេចក្ដីស្រឡាញ់ ការផ្គត់ផ្គង់ និងការលើកទឹកចិត្តរបស់អ្នកទាំងអស់គ្នាគ្មានអ្វីកាត់ថ្លៃបានឡើយ។ ក្ដីសុខរបស់អ្នកទាំងអស់គ្នា ជាក្ដីសុខរបស់ខ្ញុំដែរ។



Everyone was Made for a Mission -- ***R Warren*** **(2002)**

"My heart plans my ways: but the LORD directs my steps" (Proverbs 16:9)

PART I: SYNTHESIS

Table of Contents

Summary v

1. General Introduction 1
   1.1 Background to the study 1
   1.2 Macroinvertebrates in Asian and European rivers: a general overview 3
   1.3 Modelling techniques and applications 5
   1.4 Research problem, aims and objectives 5
2. Materials and Methods 8
   2.1 Case study in the LMB and dataset 8
      2.1.1 The LMB 8
      2.1.2 Data collection and processing 8
   2.2 Case study in Flemish rivers and dataset 10
      2.2.1 Flanders 10
      2.2.2 Data collection and processing 12
   2.3 Statistical analyses and modelling approaches 13
      2.3.1 Communities clustering and diversity measures 13
      2.3.2 Comparative analyses 14
      2.3.3 Regression and Multivariate analyses 15
      2.3.4 Model development, validation and performance 15
3. Main Results 21
   3.1 Macroinvertebrate communities and diversity patterns in the LMB 21
      3.1.1 Overall macroinvertebrate communities 21
      3.1.2 Spatial community patterns and their relationship with environmental factors 22
      3.1.3 Macroinvertebrate diversity and its relation to environmental factors 27
   3.2 Macroinvertebrate communities and diversity patterns in Flemish rivers 28
      3.2.1 Overall macroinvertebrate communities 28
      3.2.2 Spatio-temporal community composition and environmental factors 29
      3.2.3 Spatio-temporal diversity pattern and its relation to environmental factors 32
   3.3. Modelling and predicting 37
      3.3.1 Performance variation of modelling techniques applied in the LMB 37
      3.3.2 Modelling alien mollusc occurrence and their co-occurrence with native molluscs ... 39
      3.3.3 Optimizing the prediction of alien mollusc occurrence 41
4. General Discussion 43
   4.1 Overall community composition and diversity in the two river systems 43
   4.2 Spatio-temporal changes of communities and their relation to environmental factors 44
      4.2.1 The LMB 44
      4.2.2 Flemish rivers 47
   4.3 Model development, performance and predictions 51
      4.3.1 Modelling techniques and their application in the LMB 51
      4.3.2 Predicting alien species occurrence and their co-occurrence with native molluscs 52
      4.3.3 Optimizing the prediction of alien mollusc occurrence 53
5. General Conclusion and Perspectives 55
   5.1 General conclusion 55
   5.2 Implications for management and restoration 56
   5.3 Perspectives 57

References 59

Appendices 68

PART II: PUBLICATIONS

Article 1. Spatial organization of macroinvertebrate assemblages in the Lower Mekong Basin.

Sor, R., Boets, P., Chea, R., Goethals, P., Lek, S.

Limnologica (2017), 64: 20-30

Article 2. Uniqueness of sampling site contributions to the total variance of macroinvertebrate communities in the Lower Mekong River.

Sor, R., Legendre, P., Lek, S.

Ecological Indicators (revision submitted)

Article 3. Effects of species prevalence on the performance of predictive models.

Sor, R., Park, Y.S., Boets, P., Goethals, P., Lek, S.

Ecological Modelling (2017), 354: 11-19

Article 4. Spatio-temporal co-occurrence of alien and native molluscs: a modelling approach using physical-chemical predictors.

Sor, R., Boets, P., Lek, S, Goethals, P.

Aquatic Invasions (2017), 12: 147-158

Article 5. Optimizing the reliability of classification tree models in predicting alien mollusc occurrence: a hindcasting- and forecasting-based approach.

Sor, R., Boets, P., Lek, S, Goethals, P.

In preparation

Summary

Overall aims: Freshwater tropical and temperate river systems are known to support different biotic communities. In this study, I investigated benthic macroinvertebrate community composition and diversity and its spatial and temporal variation both in tropical Asian and temperate European regions. I also examined the influences of physical-chemical water quality variables on community composition, variations and diversity, and modelled the occurrence of selected species.

Locations: Tropical Asia: the Lower Mekong Basin (LMB), covering an area of 609,000 km^2; Temperate Europe: Western Europe, Flemish rivers (Belgium), covering an area of 13,787 km^2.

Materials and Methods: For the LMB, data collected from 2004 to 2008 were used, and median values of this period were analysed. For Flemish rivers, data collected from 1991 to 2010 were used. The data were divided into 4 periods: D1: 1991-1995, D2: 1996-2000, D3: 2001-2005 and D4: 2006-2010. The medians of each period were used for detailed spatial analyses. Multivariate analyses were applied to relate community composition and diversity to physical-chemical variables. Five modelling techniques namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Classification Tree (CT) were used to model the occurrence of selected species.

Main results:

Community composition variations, diversity and relationship with environmental variables

From the LMB, 299 macroinvertebrate taxa belonging to 196 genera and 90 families were identified: 131 insects, 98 molluscs, 38 crustaceans, and 32 annelids. These are the largest numbers ever reported for the LMB. Alien taxa were not assessed due to the lack of distribution data and the limited taxonomic knowledge. Insects were mostly found at the upstream parts, which were characterized by high altitude, clear water with high concentration of dissolved oxygen (DO). Molluscs, crustaceans and annelids dominated the downstream parts, which were characterized by a large surface area of watersheds, deep and wide rivers and high water temperature. These diverse compositions found at different spatial scales resulted in an increased local diversity (Shannon-Weiner, *H'*) from the upstream (*H'* = 1.9) to the downstream (*H'* = 2.7) parts, and in a great amount of total variation, or beta (β) diversity (BD_{Total} = 0.8, on a 0-to-1 scale). When computed separately for each component community (i.e. insects, molluscs, crustaceans and annelids), molluscs and insects had a higher variation and their LCBDs greatly contributed to β diversity of global communities, whereas crustaceans and

annelids had a lower variation and contributed less to the β diversity. A high degree of uniqueness in community composition in this tropical system mostly occurred along the mainstream of the rivers, which are highly associated with anthropic disturbance.

From Flemish rivers, 207 macroinvertebrate taxa belonging to 145 families were identified: 131 insects, 34 molluscs, 21 crustaceans and 21 annelids. Seventy-three alien macroinvertebrate taxa were collected. From the past (D1) to the recent period (D4), most taxa and their abundance were linked to high values of DO and low values of other water quality variables. However, Chironomidae *thummi-plumosus*, Naididae and Asellidae were found to be negatively associated with DO concentration, but positively related to the concentration of ammonium (NH_4^+), phosphate, chemical oxygen demand (COD) and nitrate (NO_3^-). Local diversity increased from D1 (H' = 1.6) to D3 (H' = 1.9), but decreased in D4 (H' = 1.6), which is due to a tremendous increase in overall abundance of macroinvertebrates in the latest period when the water quality was substantially improved. This may indicate that water quality improvement is favourable for some dominant taxa, and thus increases the homogenization of the communities and subsequently reduces the local diversity. Beta diversity of global communities was moderate, but significantly increased from the past (BD_{Total} = 0.50) to the most recent period (BD_{Total} = 0.59). When the water quality was poor (i.e. in the 1990s), a significant seasonal diversity (α and β) was observed; the diversity was low in Spring compared to Summer and Autumn. However, the seasonal difference was not found in the 2000s during which the rehabilitation program for improving water quality in most of parts of Flanders was step by step implemented. When computed separately for each component community, molluscs and insects had a high β diversity which always increased from D1 to D4, compared to β diversity of annelids (the lowest) and crustaceans which fluctuated between D1 and D4. LCBDs of the four components significantly and highly contributed to the β diversity of global communities. A high degree of uniqueness in community composition in this temperate system was more related to sites situated in the main harbour watercourses and in brackish polders, where high values of EC and pH were recorded.

Species occurrence modelling

For the LMB, the occurrence of 199 species could be predicted using LR, RF, SVM and ANN. The four modelling techniques yielded significantly different performances ($p<0.01$), of which ANN yielded the highest performance and was found to better predict the occurrence of rare species. For Flemish rivers, the occurrence of alien molluscs and their co-occurrence with native molluscs were predicted using CT. Based on field data from D1 to D4, the CT models were only able to reliably predict the "co-occurrence", but not the sole occurrence of "alien" molluscs. The co-occurrence was mainly dependent on sinuosity and a set of chemical water quality variables (e.g. NH_4^+, NO_3^-, COD, pH). When the CT models were optimized by incorporating field and cloned data (i.e. a dataset obtained by independently duplicating the field data points by *k* different individuals) via hindcasting and forecasting models, the occurrence of alien molluscs was correctly predicted with a low error rate. This result corresponds to field observations, where alien mollusc occurrence has been observed over the last two decades in Flanders.

Main conclusion and remarks: The environmental conditions of the two systems certainly favour a different macroinvertebrate community composition, and thus lead dissimilar variation. The LMB was found to support higher diversity compared to Flemish rivers. This could be due to the fact that most invertebrates from the LMB were identified to the species level, while the invertebrates from Flemish rivers were identified only to family or genus level. However, these findings indeed are not revealing the real composition and diversity for the LMB because it is very large, compared to Flemish rivers, but has been hardly studied. The number of reported taxa from the LMB is thus most likely to be underestimated. On the other hand, Flemish rivers have been extensively investigated and monitored regularly. Nevertheless, the two systems showed some similarities: molluscs and insects had a higher total variation, compared to crustaceans and annelids. Moreover, a high degree of uniqueness in community composition of the two systems mostly occurred at sites where a high level of anthropic disturbance was observed. Among the measured environmental variables, altitude, surface area of watersheds, river width and depth, Secchi depth, DO, EC and water temperature were the key factors affecting macroinvertebrate community composition and diversity in the LMB; whilst in Flemish rivers, DO, EC, pH, NH_4^+, NO_3^-, COD, phosphate and sinuosity were the key variables. These findings can provide useful information and insights which could be used to support management, conservation and restoration planning in each system.

Among modelling techniques applied in the LMB, ANN performed the best, and yielded better results when predicting the occurrence of rare species. The prediction of the occurrence of alien molluscs in Flemish rivers was successfully optimized using CT models. Whether in the past or recent periods, the results of this optimization correspond to field observations. To test for transferability, the successfully optimized models are suggested to be validated using data collected outside Flanders.

Résumé

Objectifs généraux: les systèmes fluviaux tropicaux et tempérés d'eau douce sont connus pour soutenir différentes communautés biotiques. Dans cette étude, menée dans une région d’Asie tropicale et dans une région d’Europe tempérée, j'ai étudié la composition et la diversité de la communauté des macro-invertébrés benthiques ainsi que leurs variations spatiales et temporelles. J'ai également examiné les influences des variables physico-chimiques de la qualité de l'eau sur les variations et la diversité de la composition de la communauté et j’ai modélisé l'occurrence d'espèces sélectionnées.

Localisation géographique: Asie tropicale: le bassin aval du Mékong (LMB), couvrant une superficie de 609 000 km^2; Europe tempérée: Europe occidentale, fleuves flamands (Belgique), couvrant une superficie de 13 787 km^2.

Matériel et méthodes: Pour le LMB, les données recueillies de 2004 à 2008 ont été utilisées et les valeurs médianes de cette période ont été analysées. Pour les rivières flamandes, les données collectées de 1991 à 2010 ont été utilisées. Les données ont été divisées en 4 périodes: D1: 1991-1995, D2: 1996-2000, D3: 2001-2005 et D4: 2006-2010. Les médianes de chaque période ont été utilisées pour des analyses spatiales détaillées. Des analyses multivariées ont été appliquées pour relier la composition et la diversité de la communauté aux variables physico-chimiques. Cinq techniques de modélisation, à savoir la régression logistique (LR), les Random Forest (RF), le Support Vector Machine (SVM), les réseaux de neurones artificiels (ANN) et les arbres de classification (CT) ont été utilisées pour modéliser l'occurrence desespèces sélectionnées.

Principaux résultats:

Variations de la composition des communautés, diversité et relation avec les variables environnementales

Dans le cours aval du Mékong LMB, 299 taxons de macro-invertébrés distribués dans 196 genres et 90 familles ont été identifiées; dont 131 insectes, 98 mollusques, 38 crustacés et 32 annélides. Il s‘agit du plus grand inventaire réalisé pour le bas Mékong (LMB). Les taxons exotiques n'ont pas été évalués en raison du manque de données de distribution et des connaissances taxonomiques limitées. Les insectes se trouvaient principalement dans zones amont, caractérisées par une haute altitude, une eau claire avec une forte concentration d'oxygène dissous (DO); tandis que les mollusques, les crustacés et les annélides sont majoritaires dans les parties aval, caractérisées par une grande surface des bassins hydrographiques, des rivières profondes et larges et une température élevée de l'eau. Ces compositions diverses trouvées à différentes échelles spatiales donnent lieu à une diversité locale (Shannon-Weiner, *H'*) qui s'accroit depuis les zones amont (*H'* = 1,9) vers les zones aval (*H'*= 2,7) ainsi qu'à une grande variation de biodiversité ou bêta-diversité (β) (BD_{Total} = 0,8 sur une échelle de 0 à 1). Lorsqu'ils sont calculés séparément pour chaque composante des communautés (ex. insectes, mollusques, crustacés et annélides), les mollusques et les insectes ont une variation plus élevée et leurs LCBD ont largement contribué à la diversité des communautés, alors que les crustacés et les annélides ont une variation plus faible et ont contribué moins à la bêta diversité. Un haut degré d'unicité dans la composition de la communauté de ce système tropical se produit surtout le dans les rivières, qui sont fortement associés aux perturbations anthropiques.

Dans les rivières flamandes, 207 taxons de macro-invertébrés appartenant à 145 familles ont été identifiés, dont 131 insectes, 34 mollusques, 21 crustacés et 21 annélides. Soixante-treize taxa exotiques de macro-invertébrés ont été récoltés. Dès le passé (D1) jusqu'à la période récente (D4), la plupart des taxa et leur abondance étaient liés à des valeurs élevées de DO et de faibles valeurs d'autres variables de qualité de l'eau; à l'exception du Chironomidae *thummi-plumosus*, des Naididae et des Asellidae qui se sont révélés négativement associés à la concentration de DO, mais liés positivement à la concentration d'ammonium (NH_4^+), de phosphate, de demande chimique en oxygène (COD) et de nitrate (NO_3^-). La diversité locale a augmenté de D1 (*H'* = 1,6) à D3 (*H'* = 1,9), mais a diminué en D4 (*H'* = 1,6), ce qui est dû à une augmentation énorme de l'abondance globale de macro-invertébrés dans la dernière période où la qualité de l'eau s'est considérablement améliorée. Cela peut indiquer que l'amélioration de la qualité de l'eau favorise fortement certains taxons dominants, ce qui augmente l'homogénéisation des communautés et diminue ensuite la diversité locale. La bêta diversité des communautés était modérée, mais a significativement augmenté de 1990s (BD_{Total} = 0,50) à la période la plus récente (BD_{Total} = 0,59). Lorsque la qualité de l'eau était médiocre (c'est-à-dire dans les années 1990s), une grande diversité saisonnière (α et β) a été observée; la diversité était faible au printemps par rapport à l'été et l'automne. Cependant, la différence saisonnière n'a pas été constatée dans les années 2000s au cours de laquelle le programme de réhabilitation de la qualité de l'eau dans la plupart des régions de la Flandre a été mis en place avec succès. Lorsqu'ils ont été calculés séparément pour chaque composante de la communauté, les mollusques et les insectes ont montré une grande β-diversité qui a toujours augmenté de D1 à D4, par rapport à la β-diversité des annélites (la plus faible) et des crustacés qui ont fluctué

entre D1 et D4. Les LCBDs des quatre composants ont considérablement contribué à la diversité des communautés. Un haut degré d'unicité de la composition de la communauté dans ce système tempéré était plus lié aux sites situés dans les cours d'eau du port principal et dans les polders saumâtres, où des valeurs élevées de conductivité électrique (EC) et de pH ont été enregistrées.

Modélisation d'occurrence d'espèces

Pour le LMB, l'occurrence de 199 espèces pourrait être prédite en utilisant LR, RF, SVM et ANN. Les quatre techniques de modélisation ont donné des résultats significativement différents ($p<0,01$), maisc'est l'ANN qui a donné les résultats les plus probants afin de prédire au mieux l'apparition d'espèces rares. Pour les rivières flamandes, l'apparition de mollusques exotiques et leur co-occurrence avec des mollusques indigènes ont été prédites à l'aide de CT. Sur la base des données de terrain de D1 à D4, les modèles de CT ne pouvaient prédire de manière fiable que la co-occurrence, mais pas la seule occurrence de mollusques exotiques. La co-occurrence dépendait principalement de la sinuosité et d'un ensemble de variables chimiques de qualité de l'eau (par exemple NH_4^+, NO_3^-, COD, pH). Lorsque les modèles CT ont été optimisés en incorporant des données de terrain et des données clonées (c'est-à-dire un ensemble de données obtenu par duplication indépendante des points de données de terrain par *k* individus différents) par des modèles de diffusion et de prévision, l'occurrence de mollusques exotiques a été correctement prédite avec un faible taux d'erreur. Ces résultats correspondent aux observations de terrain, où l'apparition de mollusques exotiques a été observée au cours des deux dernières décennies en Flandre.

Conclusions principales et remarques: Les conditions environnementales des deux systèmes favorisent certainement une composition différente de la communauté des macro-invertébrés et entraînent ainsi une variation différente. Nous avons constaté que le LMB possédait une plus grande diversité par rapport aux rivières flamandes. Cela pourrait être dû au fait que la plupart des invertébrés du LMB étaient identifiés au niveau de l'espèce, tandis que les invertébrés des rivières flamandes étaient identifiés uniquement à la famille ou au genre. Cependant, ces résultats ne révèlent pas vraiment la composition et la diversité réelles pour le LMB car il est très important, par rapport aux rivières flamandes, mais a été peu étudié. Le nombre de taxons du LMB est donc très probablement sous-estimé. En revanche, les rivières flamandes ont fait l'objet de suivis approfondis et ont été surveillées régulièrement. Néanmoins, les deux systèmes ont montré des similitudes; les mollusques et les insectes ont une variation totale plus élevée que les crustacés et les annélides. En outre, un degré élevé d'homogénéisation dans la composition de la communauté des deux systèmes se produit principalement dans les sites où un niveau élevé de perturbation anthropique a été observé. Parmi les variables environnementales mesurées, l'altitude, la superficie des bassins hydrographiques, la largeur et la profondeur, la profondeur de Secchi, la DO, la EC et la température de l'eau ont été les facteurs clés de la composition et de la diversité de la communauté des macro-invertébrés dans le LMB; tandis que dans les rivières flamandes, DO, EC, pH, NH_4^+, NO_3^-, COD, phosphate et sinuosité étaient les variables clés. Ces résultats peuvent fournir des informations et des idées utiles qui pourraient être utilisées pour soutenir la gestion, la conservation et la planification de la restauration dans chaque système.

Parmi les techniques de modélisation appliquées dans le LMB, l'ANN a permis d'obtenir les meilleurs résultats pour la prédiction de l'apparition d'espèces rares. La prédiction de l'apparition de mollusques exotiques dans les rivières flamandes a été réalisée avec succès à l'aide de modèles CT. Que ce soit dans les périodes passées ou récentes, les résultats de ces prédictions correspondent aux observations sur le terrain. Pour tester leur « transférabilité », les modèles les plus performants pourront être validés à l'aide de données collectées en dehors de la Flandre.

Samenvatting

Doelstelling: In deze studie werd de benthische macroinvertebraten gemeenschap en diversiteit onderzocht in een tropisch en gematigde regio en werd de spatiale en temporele variatie in kaart gebracht. Tevens werd het effect van de fysico-chemische waterkwaliteit op de samenstelling en diversiteit onderzocht.

Locatie: Tropisch Azië: het Mekong bekken met een totale oppervlakte van 609,000 km^2; Gematigd Europa: beken en rivieren in Vlaanderen (België), met een totale oppervlakte van 13,787 km^2.

Materiaal en Methoden: Voor het Mekong bekken werden data verzameld tussen 2004 en 2008 gebruikt en werd de mediaan bepaald. Voor Vlaanderen werden data verzameld tussen 1991 en 2010 gebruikt. De Vlaamse data werden in vier perioden ingedeeld, D1: 1991-1995, D2: 1996-2000, D3: 2001-2005 and D4: 2006-2010. Multivariate data analyse werd gebruikt om de samenstelling van de macroinvertebraten gemeenschap te koppelen aan de fysico-chemie. Vijf verschillende modelleertechnieken namelijk: Logistische Regressie (LR), Random Forest (RF), Support Vector Machine (SVM), Artificiële Neurale Netwerken (ANN) en Classificatiebomen (CT) werden gebruikt om de aanwezigheid van de taxa te modelleren.

Resultaten:

Samenstelling in gemeenschap, diversiteit en relatie met milieuomstandigheden

In het Mekong bekken werden 299 taxa behorende tot 90 macroinvertebraten families geïdentificeerd van de welke 131 taxa behoorden tot de insecten, 98 taxa tot de mollusken, 38 taxa tot de kreeftachtigen en 32 taxa tot de wormen. Insecten werden hoofdzakelijk teruggevonden in de stroomopwaartse delen, welke gekenmerkt worden door een relatief grote hoogte, helder water en voldoende zuurstof. Mollusken, kreeftachtigen en wormen domineerden de lager gelegen gedeelten en waren sterk geassocieerd met diepere en bredere rivieren en een hogere watertemperatuur.

De verschillende gemeenschappen die voorkwamen in de verschillende locaties resulteerde in een verhoogde lokale diversiteit (Shannon-Wiener, *H’*) van stroomopwaarts naar stroomafwaarts en een sterke variatie in de totale variatie (β diversiteit). Wanneer elke gemeenschap individueel geanalyseerd werd, zag men een hogere variatie voor mollusken en insecten en droeg hun lokale diversiteit bij aan de globale diversiteit, terwijl de variatie voor kreeftachtigen en wormen veel kleinere was en ze minder bijdroegen aan de globale biodiversiteit.

Voor Vlaamse rivieren werden er 207 taxa (behorende tot 145 macroinvertebraten families) geïdentificeerd: 131 insecten, 34 mollusken, 21 kreeftachtigen en 21 wormen. Van D1 tot D4 steeg het aantal taxa en was de abundantie voor de meeste taxa gelinkt aan een verbetering in zuurstofgehalte en een daling in nutriënten, behalve voor de tolerante taxa zoals Naididae en Chironomidae welke daalden met een verbetering in waterkwaliteit. De lokale diversiteit steeg van D1 naar D3, maar daalde in D4, wat vooral ook te wijten is aan de enorme toename in abundantie gedurende de laatste periode, welke ook gekenmerkt wordt door een verbetering in de waterkwaliteit. Dit kan er op duiden dat de verbetering in waterkwaliteit het voorkomen van bepaalde dominante soorten promoot en dus ook bijdraagt tot een homogenisatie van de gemeenschap en dus een daling kan veroorzaken van de lokale biodiversiteit. De beta-diversiteit van de macroinvertebraten gemeenschap was gemiddeld, maar steeg van D1 tot D4. In het begin van de jaren 90 van vorige eeuw, wanneer de waterkwaliteit matig tot slecht was trad er een significant seizoenale diversiteit (α en β) op, de diversiteit was eerder laag in de lente en hoger in de zomer en herfst. Echter deze seizoenale diversiteit werd niet geobserveerd begin de jaren 2000s, wanneer het herstel van onze waterlopen volop startte en zijn eerste vruchten begon af te werpen.

Wanneer de diversiteit voor elke gemeenschap apart werd berekend, hadden mollusken en insecten altijd een hoge beta-diversiteit, welke altijd steeg van D1 naar D4, dit in tegenstelling tot de beta-diversiteit van wormen en kreeftachtigen welke fluctueerde tussen D1 en D4. De diversiteit van de vier verschillende gemeenschappen droeg significant bij tot de globale diversiteit. In de brakke polderwaterlopen en in de havens werden unieke gemeenschappen vastgesteld, locaties die werden gekenmerkt door een hoge Ph en geleidbaarheid.

Soorten voorspellingsmodellen

In het Mekong bekken werden de soorten voorspeld aan de hand van LR, RF, SVM en ANN. DE verschillende modelleertechnieken gaven een verschillende uitkomst. ANN gaf het beste resultaat en was ook beter in het voorspellen van het voorkomen van weinig voorkomende soorten.

In Vlaanderen werd het voorkomen van inheemse en uitheemse mollusken voorspeld op basis van classificatiebomen. Op basis van de data waren classificatiebomen in staat om een goede voorspelling te maken van het samen voorkomen van inheemse en uitheemse mollusken, maar niet van de uitheemse mollusken alleen. Het samen voorkomen van beide soorten was voornamelijk afhankelijk van de sinuositeit en verschillende waterkwaliteitsvariabelen (e.g. NH_4^+, NO_3^-, COD, pH). Na optimalisatie van de modellen door gebruik te maken van veld data en gekloonde data (bekomen na het onafhankelijk dupliceren van de veld data door *k* verschillende individuen) en door gebruik te maken van hind- en forecasting kon het voorkomen van de mollsuken beter voorspeld worden met een lagere foutenmarge.

Conclusie en bedenkingen: De verschillende milieuomstandigheden aanwezig in de twee onderzochte systemen geven duidelijk aan dat zee en verschillende samenstelling in de macroinvertebraten gemeenschap teweeg brengen en dus ook leiden tot een verschil in variatie en diversiteit. In het Mekong bekken werd er een hogere diversiteit gevonden in vergelijking met Vlaanderen. Dit is voornamelijk te wijten aan het feit dat in het Mekong bekken macroinvertebraten tot op soort werden geïdentificeerd, terwijl in Vlaanderen slechts tot op genus of familie niveau wordt gedetermineerd. Daarenboven is ook de geografische omvang van beide systemen verschillend en dus moeilijk vergelijkbaar. Het is zo dat er in het Mekong bekken nog weinig studies rond macroinvertebraten zijn uitgevoerd en mogelijk is er dus zelfs nog een onderschatting van de werkelijk diversiteit. Desalniettemin vertoonden beide systemen toch ook enkele gelijkenissen, mollsuken en insecten vertoonden een hogere variatie in diversiteit in vergelijking met wormen en kreeftachtigen. Daarenboven werden vooral verstoorde milieu gekenmerkt door een hoge graad van uniekheid in samenstelling van de macroinvertebratengemeenschap. In het Mekong bekken waren vooral de volgende variabelen belangrijk voor de samenstelling en diversiteit van de macroinvertebraten gemeenschap: breedte en diepte van de rivier, de Secchi diepte, het zuurstofgehalte, de geleidbaarheid en de water temperatuur. In Vlaanderen waren het vooral volgende variabelen die een belangrijke rol speelden: zuurstofgehalte, geleidbaarheid, pH, ammonium, nitraat, CZV, fosforgehalte en sinuositeit. Deze informatie kan nuttig aangewend worden voor het beheer en herstel van aquatische ecosystemen.

Onder de gebruikte moelleertechnieken in het Mekong bekken, leverde ANN de beste prestatie en leverde het de beste resultaten voor het voorspellen van weinig voorkomende soorten. Het voorkomen van uitheemse mollusken kon in Vlaanderen goed voorspeld worden op basis van classificatiebomen. Om de algemene toepasbaarheid van deze modellen na te gaan wordt er voorgesteld om ook data buiten Vlaanderen te gebruiken.

1. General Introduction

1.1 Background to the study

Invertebrates, defined as organisms without backbones, are the majority of the global animal species, with an estimation of ~98% (May 1988; Chapman 2009). The invertebrates that can be seen without the aid of a microscope are referred to "macroinvertebrates", which mainly comprise insects (an estimation of ~73%), arachnids (~9%), nematodes (~7%), molluscs (~3%), crustaceans (~2%), flatworms (~1%) and annelids (~0.4%) (Chapman 2009; IUCN 2014). In freshwater ecosystems, insects, molluscs, crustaceans and annelids are the most commonly found and most studied, compared to the other groups of invertebrates (Collier and Lill 2008; Arab et al. 2009; Pérez-Quintero 2011; Szöcs et al. 2014; Sor et al. 2017a).

Macroinvertebrates are a key component of freshwater ecosystems (Palmer et al. 1997; Bogan 2008). Firstly, they are considered as ecosystem engineers (Jones et al. 1994; Chowdhury et al. 2016). For example, crustaceans and insects are responsible for regulating decomposition, shredding detritus and bioturbation. Molluscs also contribute to bioturbation, sediment formation and filtering of water, while most annelids regulate decomposition and autotrophs although some of them also promote bioturbation and sediment formation (Palmer et al. 1997). Secondly, macroinvertebrates are key networks of food chains that are important in maintaining freshwater and terrestrial food webs (Fig. 1.1). Molluscs, crabs, shrimps and other benthos (e.g. annelids and insects) are the food sources for reptiles (e.g. turtles), numerous carnivorous fish species (Chea et al. 2016), and some terrestrial predators including birds (Poulsen et al. 2004). Moreover, freshwater macroinvertebrates are generally used as bioindicators and bioassessment. For instance, the presence of mayflies, caddisflies and stoneflies (insects) can indicate good water quality conditions, while the presence of annelids, clams, pouch snails, water bugs and crayfish can indicate a moderate to high level of pollution in the water (Feld and Hering 2007; Collier and Lill 2008; Królak and Korycińska 2008; Wang et al. 2012). Due to these facts, macroinvertebrates have been used for bioassessment, which is useful to support management, restoration and conservation planning in freshwater ecosystems (Heino and Mykrä 2006; Kudthalang and Thanee 2010).

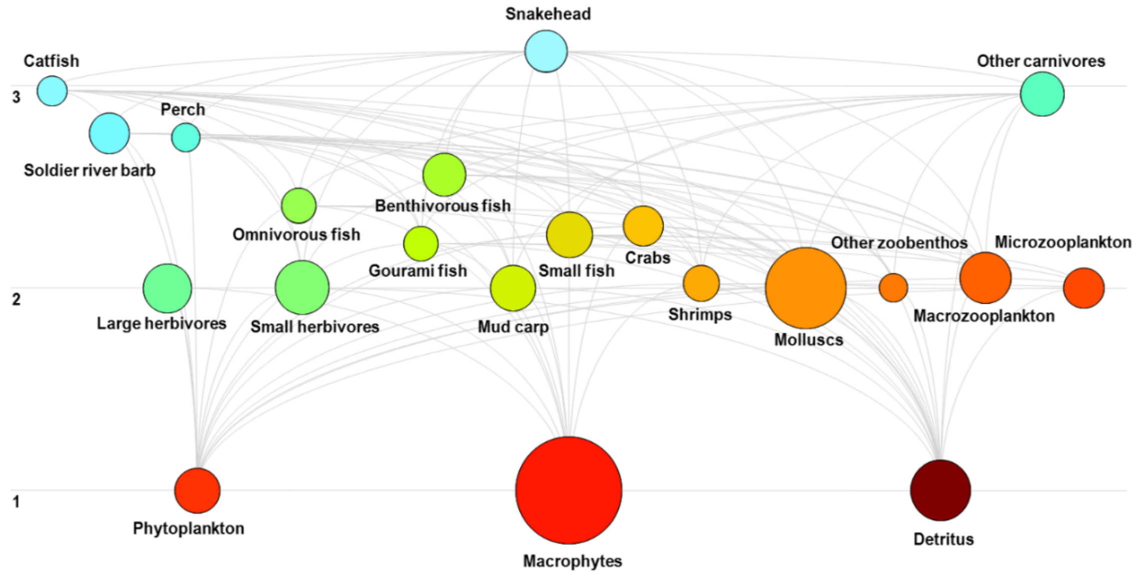


Fig. 1.1 Freshwater food web in the Tonle Sap Lake and River ecosystem, Cambodia. The size of each circle represents the biomass (tonne/km^2) of each functional group (e.g. shrimps, crabs, molluscs etc. (Chea et al. 2016).

Structure and spatial patterns of lotic macroinvertebrate communities are known to organize along a longitudinal downstream river/stream gradients (Vannote et al. 1980; Friberg et al. 2010). In the upstream parts, the communities are mainly characterized by a high abundance of insects (Collier and Lill 2008; Arab et al. 2009; Jiang et al. 2013), whereas the downstream communities are characterized by a high abundance of molluscs, crustaceans and worms (Arscott et al. 2005; Collier and Lill 2008; Pérez-Quintero 2011). For pristine or undisturbed rivers, structural and functional macroinvertebrate communities are adapted to conform to the most suitable positions, depending on the available energy flow and environmental variability, throughout the river’s continuum (RCC, Vannote et al. 1980). According to RCC, communities in the headwaters, collecting energy from decayed leaves, needles and stems, are mostly composed of shredders, collectors and less grazers/scrapers. The mid-reach, being strongly exposed to sunlight, supports more grazers/scrapers and collectors, and the lower-reach, having a low photosynthesis production in the rivers (due to high turbidity and surface film) and high energy inputs (mostly from upstream sources), is home to numerous collectors. However, in large/floodplain rivers, which receive a high level of disturbance, the RCC cannot be applied to address the biological systems. This is because differences between biotic community composition in these rivers are determined by spatial and temporal heterogeneity along the rivers (Sedell et al. 1989), and by natural and human-derived disturbances (Clarke et al. 2008; Muñoz et al. 2009).

Distribution, composition and diversity patterns of macroinvertebrates greatly vary depending on studied climatic regions (e.g. tropical vs temperate), zoogeographic regions (e.g. Palaearctic, Nearctic, Neotropical, Afrotropical, Oriental and Australasian) and geographic regions (e.g. Asia, Europe, America and Africa) (Martin et al. 2008; Yeo et al. 2008; Bogan 2008; Ferrington 2008). The variation in composition and diversity found from each climatic and zoogeographic region may reveal the different favourable environmental conditions for diverse taxonomic groups to live on (Dudgeon et al. 2006; Boulton et al. 2008). However, the ecological processes in these systems appear to be driven by more or less the same variables, e.g. drought, disturbance, nutrient concentration and trophic structure (Boulton et al. 2008; Dudgeon 2008). Nevertheless, geographic regions that lie in the tropical zone harbour a higher biodiversity, at least for most invertebrate taxa, than those lie in the southern or northern temperate zone (Sodhi et al. 2004; Boulton et al. 2008). However, stream invertebrate ecology in many parts of tropical region (e.g. South America, Africa, and Asia) remains little investigated, whereas stream invertebrates in temperate regions (e.g. North America, Europe, Australia and New Zealand) have been well studied (Dudgeon 2008; Boyero et al. 2009).

1.2 Macroinvertebrates in Asian and European rivers: a general overview

Taxonomic and ecological knowledge on freshwater macroinvertebrates in Asia, as mentioned earlier, is still limited (Boulton et al. 2008; Boyero et al. 2009). Most research in the tropical Asia is largely restricted to a few geographic regions including the Hong Kong and Peninsular Malaysia (Resh 2007; Boyero et al. 2009; Leung and Dudgeon 2011; Al-Shami et al. 2013). A wider range of studies has also been revealed from the northern subtropical Asian rivers, e.g. Yangtze and the Upper Mekong River or the so-called Lancang River in China. Research topics from these river basins include species distribution, spatio-temporal patterns and species records (Nieser et al. 2005; Shao et al. 2008; Qi et al. 2012). However, most of the research findings are reported in Chinese and are not publically available; only a few are accessible, e.g. the benthic macroinvertebrates as indicators of ecological status in Yangtze River (Pan et al. 2013) and the seasonal variability of metazooplankton (including crustaceans) communities and new mollusc species records from the Upper Mekong Basin (Du et al. 2011; Wu et al. 2014). For the Lower Mekong Basin, more investigations have been recently conducted in Thai streams (Boonsoong et al. 2010; Kudthalang and Thanee 2010; Phaphong and Sangpradub 2012; David and Boonsoong 2014) and recently also some Philippine streams have been investigated (Tampus et al. 2012; Sinco et al. 2014; Fajardo et al. 2015; Magbanua et al. 2015, Forio et al. 2017). Most of these studies are related to species diversity, description of new species and using benthos to assess water quality in river systems (Parnrong et al. 2002; Sangpradub et al. 2002; Flores and Zafaralla 2012). Macroinvertebrates from other geographic areas including Myanmar, Laos, Cambodia and Vietnam, which mainly share the Lower Mekong Basin, remain very scarce.

On the contrary, knowledge on macroinvertebrates from river systems in Europe has been extensively studied (Boyero et al. 2009). Since the adoption of the European Water Framework Directive (WFD) (European Commission 2000), freshwater macroinvertebrates have become the central focus (Pollard and Huxham 1998; Hering et al. 2010). Macroinvertebrates from hundreds of streams have been studied and used to assess water quality (Buffagni et al. 2001; Verdonschot and Nijboer 2004). Within 10 years of the implementation, ~1,900 papers resulted from research projects associated with WFD (Hering et al. 2010). This results in a very well documented knowledge on freshwater macroinvertebrates and their ecological applications for Europe. Furthermore, a diverse assessment methods have been developed (Birk et al. 2012), some of which have applied a predictive modelling framework that is based on macroinvertebrates or use environmental variables to predict future distribution, occurrence and abundance of particular taxa (Goethals et al. 2007; Everaert et al. 2013; Boets et al. 2015).

When macroinvertebrate composition and diversity are related to measured environmental variables, key factors driving spatio-temporal changes have been known to be more or less the same regardless of geographic or climatic regions. For instance, macroinvertebrate communities in river basins from southern China (Pearl, Yangtze and Qiangtang rivers), from northern Portugal (the Olo, Corgo, Pinhao and Tua rivers) and from Susquehanna River (New York, North America) have been reported to be influenced by land use types including anthropogenic disturbance (Allan 2004; Bruns 2005; Cortes et al. 2011; Cortes et al. 2013). Another example can be found from European Mediterranean (Evrotas River, Greece) and Asian streams (Peninsular Malaysia) that stream size (e.g. width and depth), dissolved oxygen and pH were the key factors influencing macroinvertebrate composition and variation (Al-Shami et al. 2013; Salmah et al. 2014; Karaouzas and Płóciennik 2016). These indicate that similar ecological processes can be expected from different ecological systems.

1.3 Modelling techniques and applications

Various modelling techniques have been widely and increasingly implemented in ecological systems (Lek et al. 1996; Park et al. 2003; Schröder et al. 2007; Lencioni et al. 2007; Guo et al. 2015). The techniques applied are generally used to explain and predict the relationship between the occurrence or abundance of studied species and environmental variables (Goethals et al. 2007; Boets et al. 2013). Utilization of modelling techniques to combine both explaining and predicting such relationships is also commonly applied (Roura-Pascual et al. 2009; Call et al. 2016). Applications of predictive models have provided knowledge and improved the understanding of the ecology and behaviour of studied taxa, which could be used to support decision making, management and conservation planning. For instance, many previous studies have used predictive models to predict the occurrence and distributional areas of plants, herbs, macroinvertebrates and fish (Thuiller et al. 2005; Roura-Pascual et al. 2009; Vicente et al. 2011; Boets et al. 2013; Chen et al. 2015; Guo et al. 2015).

However, the application of predictive models has been suggested to be carefully taken into account because they can have a wide variation in performance (Segurado and Araujo 2004; Elith et al. 2006; Guisan et al. 2007). Some models even yield contrasting predictions of habitat suitability (e.g. Guisan et al. 2007; Evangelista et al. 2008; Roura-Pascual et al. 2009). Furthermore, predictive models are sensitive to parameterization and selection criteria during the modelling process (Araújo and Guisan 2006; Elith et al. 2006), and thus can result in an uncertainty of current or past/future projections of species distributions (Svenning et al. 2008; Buisson et al. 2010). Due to this fact, when calibrating and validating predictive models, carefully taking into account the data characteristics (e.g. sample size, species prevalence or environmental predictors), parameterization and selection criteria are usually recommended (Luoto et al. 2006; Dormann et al. 2008).

1.4 Research problem, aims and objectives

The Lower Mekong Basin (LMB), which includes portions of Thailand, Laos, Cambodia and Vietnam, is characterized by a long and large floodplain (Eastham et al. 2008) and is known for its high biodiversity (Sodhi et al. 2004). However, the knowledge of macroinvertebrates in the LMB is poorly investigated. Given that this river basin is being impacted by various anthropogenic disturbances such as agricultural activities, aquaculture, urbanization and mining (Sodhi et al. 2004; Nhan et al. 2007; Köhler et al. 2012), there is an urgent need to study the patterns of spatial organization, community structure and variations of macroinvertebrates in this basin and their relation to environmental factors. Up to date, only a few studies (except for those conducted in Thailand) have been published on the basin, e.g. community structure and composition of littoral invertebrates in the Mekong delta (Wilby et al. 2006) and the diversity and distribution of crustaceans and molluscs in the Indo-Burma region (Cumberlidge et al. 2011; Köhler et al. 2012). Yet, no attempt has been made to examine the large spatial patterns, community structures, variations (i.e. β diversity) of macroinvertebrate communities and their relation to key environmental variables nor the application of predictive modelling in this hardly studied basin.

On the other hand, river systems in Europe as well as in Flanders suffered from severe water quality degradation in previous decades (e.g. from 1980s to 1990s). During these periods, some native species were reported to disappear (Bernauer and Jansen 2006) and only those that were able to withstand the water quality degradation remained. At the same time, most European river systems have been exposed to a number of alien macroinvertebrate species (Leuven et al. 2009; Boets et al. 2016). From the 2000s until now, the water quality of European rivers has been greatly improved. This water quality improvement does not only promote the occurrence and abundance of native species, but also favours the alien species to spread widely, which consequently may lead to changes in community composition. As such, investigation spatio-temporal changes in community composition, variations and predicting the occurrence of alien species across Flemish rivers, which have been poorly studied, will provide insights into the ecology of overall communities and of studied alien species. Results from this investigation can be used to support management and conservation planning.

The aims and specific objective (or questions) of the present study are:

Aim 1. Investigating general patterns of macroinvertebrate communities and their relation to environmental variables in the two systems, i.e. the LMB and Flemish rivers.

1.a. Investigating patterns of spatio-temporal variation in macroinvertebrate assemblages/communities.

1.b. Analysing the variability of macroinvertebrate composition among the assemblages/communities, and determining key indicator/important taxa (the most representative taxa/taxa with high among-site variance).

1.c. Identifying the important environmental variables that are associated with the particular macroinvertebrate assemblages/communities.

Aim 2. Determining the total variation in macroinvertebrate communities (i.e. total β diversity) and the key determinants in the two systems.

2.a. Is there a moderate or a large amount of total β diversity?

2.b. What are the taxa that contribute most to the total β diversity?

2.c. What are the environmental conditions and component communities (e.g. annelids, crustaceans, insects and molluscs) that significantly influence the total β diversity?

Aim 3. Predicting macroinvertebrate species occurrence and analysing the performance of modelling techniques applied in the LMB.

3.a. Predicting the occurrence of macroinvertebrate species and comparing the performance of the applied techniques based on a complete prevalence range (i.e. 0.0-1.0), and different prevalence ranges (i.e. at a 0.1 interval).

3.b. Analysing how the species prevalence affects the behaviour of modelling techniques' performance.

Aim 4. Predicting the occurrence of alien species and their co-existence with native species and identifying the key determining variables in Flemish rivers over the past two decades (1991-2010).

4.a. Identifying key determining physical-chemical variables associated with the occurrence of alien species (i.e. alien molluscs) and with the co-occurrence of alien and native species, using a classification tree modelling technique.

4.b. Optimising the reliability of classification tree models in predicting alien mollusc occurrence.

2. Materials and Methods

2.1 Case study in the LMB and dataset

2.1.1 The LMB

The Mekong River Basin is divided into the Upper Mekong Basin (UMB) and the Lower Mekong Basin (LMB). The UMB on the Tibetan plateau in China is composed of narrow, deep gorges and small, short tributaries, whereas the LMB stretches from Yunnan province in South China to the delta in Vietnam and it covers approximately 70% of the total length of the whole basin (Eastham et al. 2008). The LMB consists of a large floodplain and long, broad tributaries and it drains more than 76% of the Mekong basin. The climate of the LMB is dominated by a tropical monsoon rainfall system, which is characterized by a dry (November – April) and a wet (May – October) season generated by the northeast monsoon and the south-west monsoon, respectively. The most intensive rainfall falls from July to September, while the lowest precipitation is observed between January and April (Adamson et al. 2009). The annual rainfall of the LMB varies from 1,000 – 1,600 mm in the driest regions to 2,000 – 3,000 mm in the wettest regions (Hoanh et al. 2003). A higher precipitation is found in the eastern mountainous regions of Laos and in northeast Thailand (Eastham et al. 2008).

The largest floodplain water body of the LMB is the Tonle Sap Lake (TSL) in Cambodia (Adamson et al., 2009), which is the largest freshwater lake in Southeast Asia (Sarkkula et al. 2003). The TSL is connected to the Mekong through the Tonle Sap River, and thus creating an exceptional hydrological cycle. In the wet season, the TSL receives excess water from the Mekong River and expands its surface area from 2,500 km^2 to 15,000 km^2. In the dry season when the rain ceases and water levels drop in the Mekong, a reverse flow occurs; the drained water from the TSL flows to the Mekong delta (Arias et al. 2011). The Mekong delta is characterized by a number of man-made canals, which are mostly used for domestic and agricultural activities (Kummu et al. 2008).

2.1.2 Data collection and processing

Benthic macroinvertebrates were sampled at 60 sampling sites along the main channel of the LMB and its tributaries by the Mekong River Commission (MRC) (Fig. 2.1). This sampling was carried out once a year in March during the dry season from 2004 to 2008. At each sampling site, macroinvertebrates were sampled from three locations in the benthic zone: near the left and right banks, and in the middle of the rivers. At each location, a minimum of three samples (where inter-sample variability is low, e.g. tributaries) to a maximum of five samples (where inter-sample variability is higher, e.g. the main channel and the delta) were collected using a Petersen grab sampler which has a sampling area of 0.025 m^2. With the grab sampler, four sub-samples were taken and pooled to give a single sampling unit covering a total area of 0.1 m^2. In total, between nine (3 samples × 3 locations) and fifteen (5 samples × 3 locations) pooled samples were collected at each sampling site. Each pooled sample was rinsed using a sieve (0.3 mm mesh size). In the field, the samples were sorted and then preserved by adding 10% formaldehyde to obtain a final concentration of about 5%. In the laboratory, they were identified to the lowest taxonomic level possible and counted using a compound microscope (40 – 1,200 magnification) or a dissecting microscope (16 – 56 magnification). Macroinvertebrate abundance data per sampling unit was averaged across all samples (between 9 and 15 samples) collected from each sampling site.

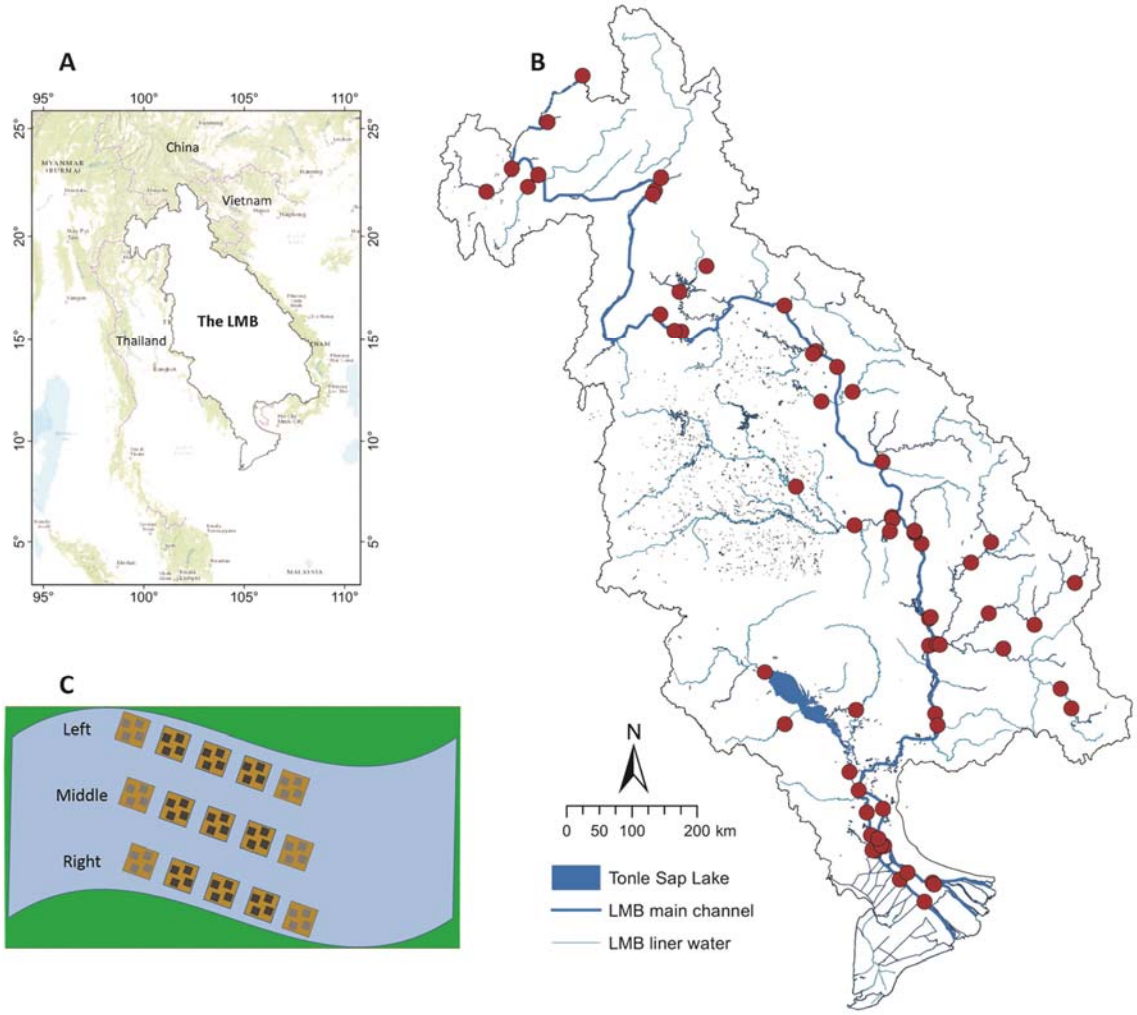


Fig. 2.1 The Lower Mekong Basin (LMB, A) and macroinvertebrate sampling sites (shaded dots, B). Sub-samples and replicates were taken at each sampling site as illustrated in C.

At the sampling site, geographic coordinates and altitude were determined with a GPS (Garmin GPS 12XL). River width was measured in the field using a Newcon Optik LRB 7x50 laser rangefinder. Other physical-chemical variables were measured at the three locations where macroinvertebrates were sampled. River depth was measured using a line metre. With a handheld water quality probe (YSI 556MP5), water temperature, dissolved oxygen, pH and electrical conductivity were measured at the surface (0.1-0.5 m) and at a depth of 3.5 m or at a maximum depth of the river (wherever less than 3.5 m) and then the average value was recorded for each location. Water transparency was measured with a Secchi disc by lowering it into the water and recording the depth at which it was no longer visible. The physical-chemical data of each sampling site was the averaged value across the three sampling locations. Distance from the sea and the surface area of watersheds drained at each sampling site was determined using a Geographic Information System (ArcGIS 10.0, ESRI). Geographic data (ArcGIS shapefiles) about the LMB (river networks, basin boundaries, land covers, and subcatchments derived from topographical maps) was provided by the MRC.

In total, 108 samples were collected from the 60 sampling sites. In 2008, 3 sampling sites were sampled further away from their original sampling coordinates, and thus were considered as different sampling sites (see Appendix T1). Therefore, a total of 63 sampling sites were taken into account in the analyses. Because of unequal sampling efforts (i.e. unequal and different number of samples at each site during the 5-year sampling period) and missing values of environmental variables, we used median values from the collected data to represent each site in the analyses, as suggested by McCluskey and Lalkhen (2007). These median values were used in all of the analyses corresponding to the case study of the LMB.

2.2 Case study in Flemish rivers and dataset

2.2.1 Flanders

Flanders (northern Belgium) is located in Northwest Europe and its Northwestern part is bordered by the North Sea (Fig. 2.2A). Flanders has a total area of 13,522 km^2, and is considered as one of the most densely populated regions in Europe (477 inhabitants/km^2 in 2015, https://en.wikipedia.org/wiki/). Flanders is classified as a lowland area, which is divided into different rivers basins (Fig. 2.2B). This region is influenced by a temperate oceanic climate, as same as most of northwestern European countries are (e.g. UK, France, Luxembourg, Netherland and Denmark) (Peel et al. 2007). Flanders has a dense watercourse network including navigable canals. Agriculture, industry and residential areas are the main land use types of Flanders and its landscape is characterized by highly fragmented and complex mosaic of land use types (Poelmans and Van Rompaey 2009). This fragmentation and complexity may have put a high pressure on habitat quality and biodiversity in Flanders.

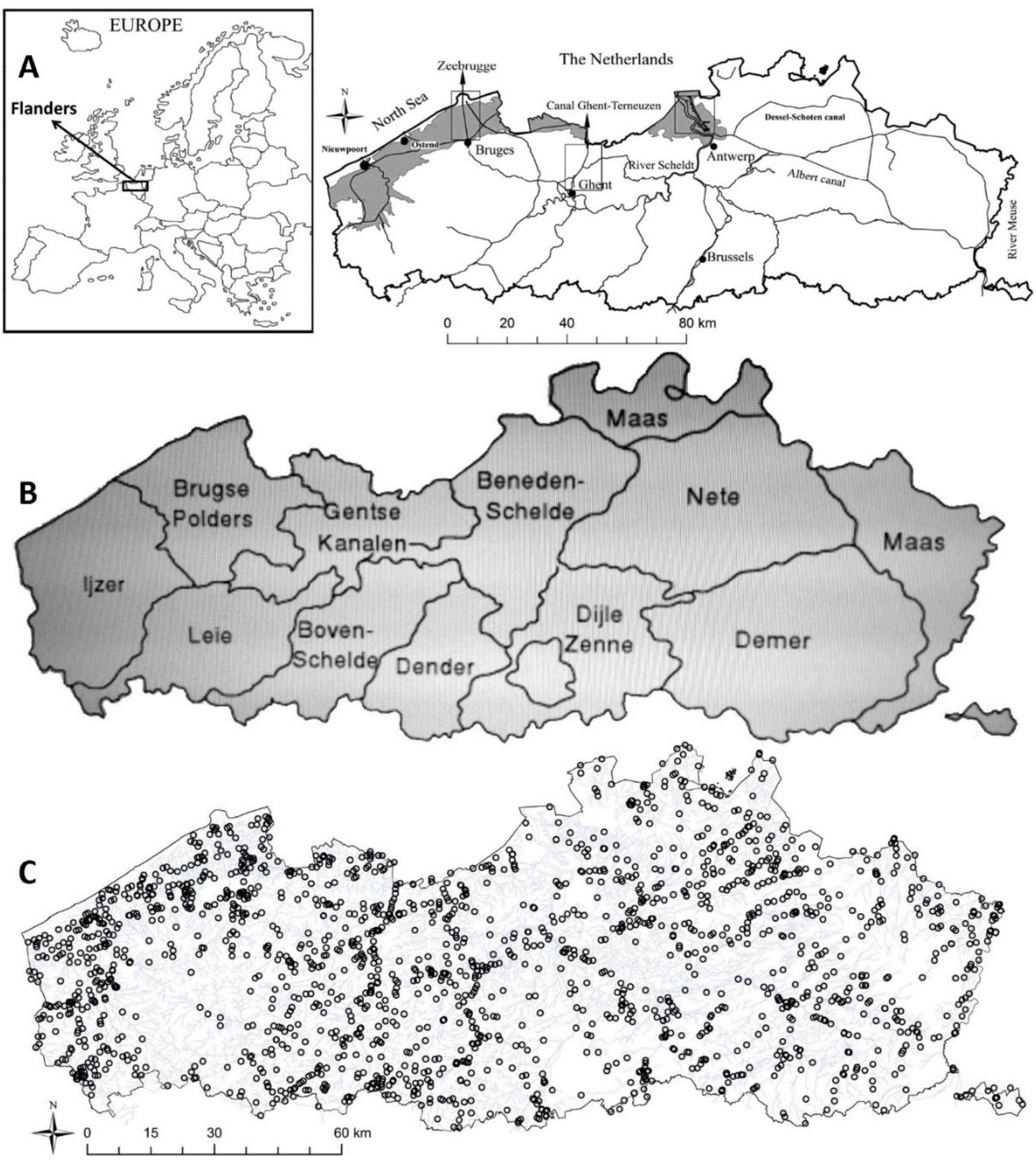


Fig. 2.2 Map of Flanders indicating: (A) the most important watercourses and geographic locations, the polder area (grey) and the three main harbours indicated by rectangles (Boets et al. 2016), (B) different river basins (van Griensven and Vandenberghe 2006) and (C) monitoring sites between 1991-2010, which were used in the present study.

2.2.2 Data collection and processing

The Flemish Environment Agency (VMM) has collected biological and environmental data in Flanders since 1989. The monitoring sites include all types of watercourses from all river basins. Every three year from the beginning, a fixed set of sampling locations was sampled. Most of the sampling locations were only sporadically sampled, and thus results in a large dataset of more than 11,000 biological samples collected at more than 2500 sites spread over different water bodies (Fig. 2.2C). In this monitoring program, the sampling protocol was entirely based on the method as described by Gabriels et al. (2010). Macroinvertebrates were collected using a standard handnet, which is made of a metal frame (0.2 m by 0.3 m) to which a conical net is attached with a mesh size of 300 µm. The kick sampling was made along the watercourses at a stretch of approximately 10-20 m. Each sample was collected for three minutes for small watercourses (less than 2 m wide) or five minutes for larger rivers. At sampling sites where the kick sampling method was not possible, artificial substrates were used. Three replicates of artificial substrates, which consisted of polypropylene nets (5 litres) filled with bricks of different sizes, were left in the water for a period of at least three weeks after which they were retrieved. Leaving this period enables species to colonize the substrates. The different sampling efforts of the two sampling approaches (the kick and artificial substrate sampling) may have repercussion on the diversity of sampled invertebrates. However, according to Gabriels et al. (2010), the two approaches are standardized semi-quantitative methods and are similar in terms of sampled macroinvertebrate abundance. In the laboratory, macroinvertebrates in the VMM database were identified to the level (family or genus) needed for the calculation of the biotic water quality index.

Electrical conductivity (EC), pH and dissolved oxygen (DO) were measured in the field with a hand-held probe (Cond 315i, oxi 330, wtw, Germany and 826 pH mobile, Metrohm, Switzerland). All additional chemical variables, i.e. ammonium (NH_4^+), chemical oxygen demand (COD), biological oxygen demand (BOD), total phosphorus (Pt), nitrate (NO_3^-), nitrite (NO_2^-), Kjeldahl nitrogen, orthophosphate (oPO_4), were retrieved from the monitoring dataset compiled by the VMM and which is online accessible (www.vmm.be). Nutrient analysis was performed spectrophotometrically in accordance to ISO 17025. GIS software (version 9.3.1) applied on the Flemish Hydrographic Atlas was used to determine the slope and the sinuosity of a watercourse at a different height in between two points (1000 m apart) and on a stretch of 100 m, respectively.

Data from 1991 to 2010 was used for the analyses. Based on the preliminary data mining, the overall communities had a temporal change (especially a somewhat different community composition for the late 2000s) which could be grouped based on a five-year interval. Therefore, the data were then divided into 4 periods. Each period consisted of samples from a five-year sampling effort (i.e. D1: 1991-1995, D2: 1996-2000, D3: 2001-2005 and D4: 2006-2010). This division can provide useful information on changes in community composition for each period. To analyse spatial variation in the community composition, the median values were used to represent each site for each period. This is because, as mentioned earlier, only a fixed set of sampling locations was sampled regularly whereas most of the other sampling locations were sporadically sampled. For the modelling purposes (see the detailed in the "Modelling" section below), all collected samples were used.

2.3 Statistical analyses and modelling approaches

I analysed only four groups of macroinvertebrates in this study. This is because they are the most commonly studied animals and are generally used as bioindicators and assessment in freshwater ecosystems (Feld and Hering2007; Collier and Lill 2008; Wang et al. 2012). The four groups included annelids, crustaceans, insects and molluscs. These four groups were designated as component communities in following paragraphs onwards. All the applied statistical analyses and modelling approaches were performed using functions of packages in the R language program (R Core Team 2013).

2.3.1 Communities clustering and diversity measures

Samples were clustered based on the Bray-Curtis dissimilarity of macroinvertebrate abundance data by using Ward's hierarchical method. The Bray-Curtis dissimilarity distance (Legendre and Legendre 2012) between the macroinvertebrate samples was calculated using the Hellinger transformation in the package *vegan* of R (Rao 1995).

The macroinvertebrate indicator taxa in each assemblage were determined using the Indicator Value (IndVal, Dufrene and Legendre 1997) with the package *labdsv* of R (Roberts 2013). The Indicator Value of a taxon is an index ranging from 0 to 1, indicating the least to most important taxa occurring in a group of sites. A value of 1 is obtained when every individual of the taxon is found only in the group and when it occurs at all sites of that group. A high number of taxa with significant Indicator Values may provide information on the habitat they prefer to share. Taxa having Indicator Values with a p-value ≤0.01 were retained as the most important taxa representing a given assemblage (consisting of a group of sites).

Macroinvertebrate richness, abundance and Shannon-Wiener diversity (*H'*) were calculated for each sampled site, cluster and group. To quantify beta (β) diversity, the community composition data were first Hellinger-transformed (Legendre and Gallagher 2001; Legendre and Legendre 2012). For Hellinger-transformed data, the total variance, or total β diversity (BD_{Total}), of a community composition data table is an index between 0 and 1, and it can be partitioned into local contribution (LCBD) and species contribution (SCBD) indices. An LCBD value is an index showing the degree of uniqueness in taxonomic composition in each site, computed as the relative contribution of a site to BD_{Total}, so that the LCBD indices sum to 1, whereas an SCBD index shows the relative degree of variation of a taxon across all sites. The BD_{Total}, LCBD and SCBD indices were computed using the function "beta.div" available in the *adespatial* package in R (Dray et al. 2016). The Hellinger transformation was used because the corresponding Hellinger distance is one of the dissimilarity functions admissible for β diversity analyses (Legendre and Gallagher 2001; Legendre and De Cáceres 2013); it does not give high weights to the rare species. In addition to LCBD, Hellinger-transformed data also allow researchers to compute SCBD indices; this is not allowed by most other admissible dissimilarity functions (Legendre and Gallagher 2001; Legendre and De Cáceres 2013). SCBD indices that were higher than the mean of SCBD values identified the taxa that were the most important contributors to BD_{Total}. In the following paragraphs, BD_{Total}, LCBD and SCBD designate the indices of the global macroinvertebrate communities, whereas BD_{ATotal}, BD_{CTotal}, BD_{MTotal}, BD_{ITotal}, and $LCBD_{A}$, $LCBD_{C}$, $LCBD_{M}$ and $LCBD_{I}$ designate the BD_{Total} and LCBD indices for annelid, crustacean, mollusc and insect communities, respectively.

2.3.2 Comparative analyses

Descriptive statistics were used to describe and summarize the information of the collected data. These included minima, maxima, mean, range, standard deviation (sd), sample size, and percentage. In most cases, mean and standard errors were used to indicate significant differences in macroinvertebrate composition, environmental conditions and model performances between/among groups. Where applicable and appropriate, a one-way ANOVA or a Kruskal-Wallis test was used to test for significant differences between/among unmatched groups (≥3 groups). One-way ANOVA was applied when residuals of the models were normal (Shapiro-Wilk test, $p > 0.05$, and homoscedastic (Bartlett's test, $p > 0.05$)); otherwise, the non-parametric test (Kruskal-Wallis) was used. For matched groups (dependent samples), a multi-factor ANOVA and a Friedman test were used when the data was normal and not-normal distributed, respectively.

2.3.3 Regression and Multivariate analyses

Simple and multiple regression models were used to access the influence of independent variables (e.g. environments) on response variables (e.g. communities, diversity measures and indices). To identify the strength of the regression models, the stepwise selection with the Akaike Information Criterion (AIC) was applied. The models having the lowest AIC and highest adjusted R^2 were considered to have the strongest influence on the response variables.

Linear Discriminant Analysis (LDA) was performed, using the package *ade4* of R (Chessel 2006), to assess which measured environmental variables best accounted for the differences among the macroinvertebrate assemblages grouped by the hierarchical clustering. Before performing the LDA, environmental variables were tested for multivariate homogeneity of within-group covariance (Borcard et al. 2011). The contribution of each variable to the discrimination among assemblages was represented by the standardized factorial coefficient, projected as an arrow on the LDA plot.

With a complex community data, as in the case of Flemish river data, Redundancy Analysis (RDA) was conducted on the Hellinger-transformed abundance data and environmental factors. RDA is powerful tool for the analysis of community composition data tables (Legendre and Legendre 2012). The RDA model was first tested at global scale to detect for its significance, and afterwards, the forward selection method was carried out in order to select the most importing factors associating with the community composition.

2.3.4 Model development, validation and performance

Five commonly used modelling techniques were applied in this study. They included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Classification Tree (CT). For detailed information on the above mentioned techniques as well as for practical examples of the used methods, see Van Echelpoel et al. (2015), Sor et al. (2017b) and Article 4 in Part II: Publications.

2.3.4.1 Species occurrence prediction in the LMB

Macroinvertebrate species occurrence was predicted using four modelling techniques: LR, RF, SVM and ANN. The presence/absence of each species was used as the response variable and the measured environmental variables such as altitude, river width, river depth, distance from the sea, water temperature, DO, EC, pH and Secchi depth were used as the input predictors. For every prediction, species that occurred only in one instance (a prevalence of ~0.02) were removed from the data. This is because Leave-One-Out cross-validation (LOO) was used to validate the models, due to small sample size. With LOO, it is not feasible to split data into training and validation sets for species that have only one occurrence instance.

To evaluate the four modelling techniques, Cohen's Kappa Statistic (Kappa), area under receiver operating characteristic curve (AUC), and error rate (ER) of the prediction were used. The overall performance differences among the four modelling techniques were assessed using a Friedman test (nonnormal distribution). P-values ≤0.05 were considered to indicate significant differences. Each performance measure of the four modelling techniques was then regressed against the species prevalence using two types of models: a linear and a quadratic model. Mean performance values and standard errors were compared to identify which technique better performed based on the complete prevalence and based on different prevalence ranges.

2.3.4.2 Species occurrence prediction in Flemish rivers

A CT model was used to predict the occurrence of alien and native molluscs and their co-occurrence, and to identify the determining physical-chemical variables. The CT was implemented in this system because it is well suited for analyses of complex ecological data (De'ath and Fabricius 2000). Moreover, it is relatively simple to implement, easy to interpret, and it tolerates missing values during both the training and testing cases (Therneau and Atkinson 1997; De'ath and Fabricius 2000). The prediction was made for each period. The response variable of each sampling site of each period was categorized as: a "native" site (i.e. a site having only native molluscs present), an "alien" site (i.e. a site having only alien molluscs present) and a "co-occurrence" site (i.e. a site having both alien and native molluscs present). Due to a limited number of occurrence instances of most alien mollusc genera, all alien genera were merged to form one categorical variable. This provided a higher number of instances for the predictive models and thus a better and more robust development of the model. The results of these predictions could reveal common environmental conditions that most of the alien molluscs prefer. In the same way, all native genera were also merged to form one categorical variable. Physical and chemical water quality variables (i.e. BOD and Kjeldahl nitrogen) that had missing values for more than 5% of the total samples were removed from the analyses. Therefore, each period consisted of one response categorical variable (native/alien/co-occurrence) and 11 predictor variables. The summary of the physical and chemical water quality variables and of the response variable is shown in Table 2.1. During the model development, all CT trees were initially pruned by setting a complex parameter at cp =0.05. Where the tree had only a root, the cp was lowered to a level (e.g. cp =0.04, 0.03, 0.02, 0.01) that at least two terminal nodes were produced. In practice, the first few splits mostly provide a very informative division of the data (Therneau and Atkinson 1997). These criteria were set in order to make the trees easily interpretable and comparable in terms of the number of variables and complexity.

Table 2.1 Mean value (and standard deviation) for environmental predictors, and occurrence instances of each class of the response variable. NH_4^+: Ammonium, COD: Chemical Oxygen Demand, Pt: Total Phosphorus, EC: Electrical Conductivity, NO_3^-: Nitrate, NO_2^-: Nitrite, oPO_4: Orthophosphate, DO: Dissolved Oxygen. The number of sampled sites for each period is shown in square brackets.

		Period			
		1991-1995	1996-2000	2001-2005	2006-2010
Variable	Unit	[509]	[991]	[1524]	[1250]
NH_4^+	mg/L	2.6 (4.8)	1.9 (3.4)	1.6 (2.8)	2.3 (5.1)
COD	mg/L	55 (43)	39 (42)	34 (34)	36 (41)
Pt	mg/L	1.0 (1.4)	0.9 (1.4)	0.9 (2.3)	0.8 (1.0)
EC	µS/cm	1320 (2618)	987 (1149)	998 (1509)	921 (949)
NO_3^-	mg/L	3.3 (4.1)	4.0 (4.7)	3.6 (3.7)	3.0 (3)
NO_2^-	mg/L	0.2 (0.3)	0.2 (0.3)	0.2 (0.2)	0.2 (0.2)
oPO_4	mg/L	0.6 (0.9)	0.5 (1.1)	0.4 (0.7)	0.5 (0.8)
pH		7.5 (0.6)	7.6 (0.5)	7.7 (0.4)	7.6 (0.4)
DO	mg/L	7.5 (3.5)	6.8 (3)	6.9 (3.1)	6.6 (2.9)
Sinuosity		1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)
Slope	m/1000m	1.0 (1.4)	1.6 (2.8)	2.0 (3.6)	1.7 (2.8)
Response	class				
Co-occurrence		*66*	*146*	*567*	*875*
Alien		*6*	*15*	*63*	*133*
Native		*863*	*1860*	*2259*	*842*
Total instances		935	2021	2889	1850

For each period, a three-fold cross-validation was used to train and validate the models. To build reliable models and to avoid misidentifying the key variables determining each class of the response variable, 3 replications of the three-fold cross validation was made. For each 3-fold cross-validation, the data was shuffled and randomly split into three subsets; two subsets were used for training and one subset for validation. For the second and third replication, I re-shuffled and randomly split it into new training and validation sets following the same procedures. From each training and validation set, a model was built and in this way, a performance value and the importance of each variable (in percentage) of nine different models (3 models of each three-fold cross validation × 3 replications) were calculated. A mean performance value, obtained from the nine models, was used as a final criterion for model evaluation. Cohen's Kappa Statistic (Kappa) and Correctly Classified Instances (CCI) were used to evaluate the model performance. The importance of each variable determining the preferred environmental conditions of each class (native/alien/co-occurrence) was averaged across the nine models. To identify which variables significantly determine the preferred conditions of each class, the importance of each variable was compared based on the standard error. The same procedures and criteria were applied for the modelling of data of each period.

2.3.4.3 Optimizing alien mollusc occurrence prediction

Due to the small number of occurrences, the CT models were not able to predict sites where only alien molluscs occurred (see Results section 3.3.2). Subsequently, the CT models were optimized by combining the field and cloned data, and testing two types of resampling approaches (stratified and random) under same and different parameterization settings.

The combined data was prepared as follows. First, a cloned dataset was developed by independently duplicating the original field data of the less frequently occurring classes (mostly the "alien" and "co-occurrence" class) by *k* different individuals. Then the cloned data was combined with the original field data to obtain the same sample size for each class (Fig. 2.3A). One may argue to use an equal-stratified dataset of field observations by randomly selecting a number of observations of the more frequently occurring classes in an equal number as there are of the less occurring class. By following this routine, valuable information of many of the unselected field observations will be excluded. That would in this case lead to a small sample size for the model development, which consequently leads to a less reliable and less robust models. On the contrary, using the cloned data to combine with all field observations yields an equal data distribution and sufficient samples for each period. Incorporating the cloned data into the models will increase the maximum likelihood of the prior class distribution, making the models more robust (Lele et al. 2007).

Model optimization was carried out in two scenarios: a hindcasting and forecasting. For the hindcasting, only data period D4 was used to calibrate the models. Then the calibrated models were used to hindcast (validate) the response variable based on the environmental predictors of the data from period D4, D3, D2 and D1. For the forecasting, only data period D1 were employed to calibrate the models. Then the calibrated models were used to forecast (validate) the response variable based on the environmental predictors of the data from period D1, D2, D3 and D4. During the pruning phase, the complex parameter of all calibrated models was set in the same way for both scenarios, following the procedure described in an earlier section.

The calibration and validation process was based on a three-fold cross-validation (CV), following the procedure described earlier. When splitting the dataset for the 3 fold CV, a random and a stratified resampling approach were applied. The random resampling was made by shuffling and randomly splitting the whole dataset into three folds. From this approach, the sample size of each class was not equally stratified across the three folds. To allow a reliable error estimation of the hindcasting and forecasting models for each period, three replicates of the 3 fold CV were made by reshuffling and randomly splitting the whole dataset again into 3 new folds. For the stratified resampling, the three classes of the response variable and the environmental predictors of the whole dataset were first allocated into three separated data subsets (DS, i.e. DS1, DS2, DS3). Each DS corresponds to the data of each class of the response variable and its corresponding environmental predictors. Then, each DS was randomly divided

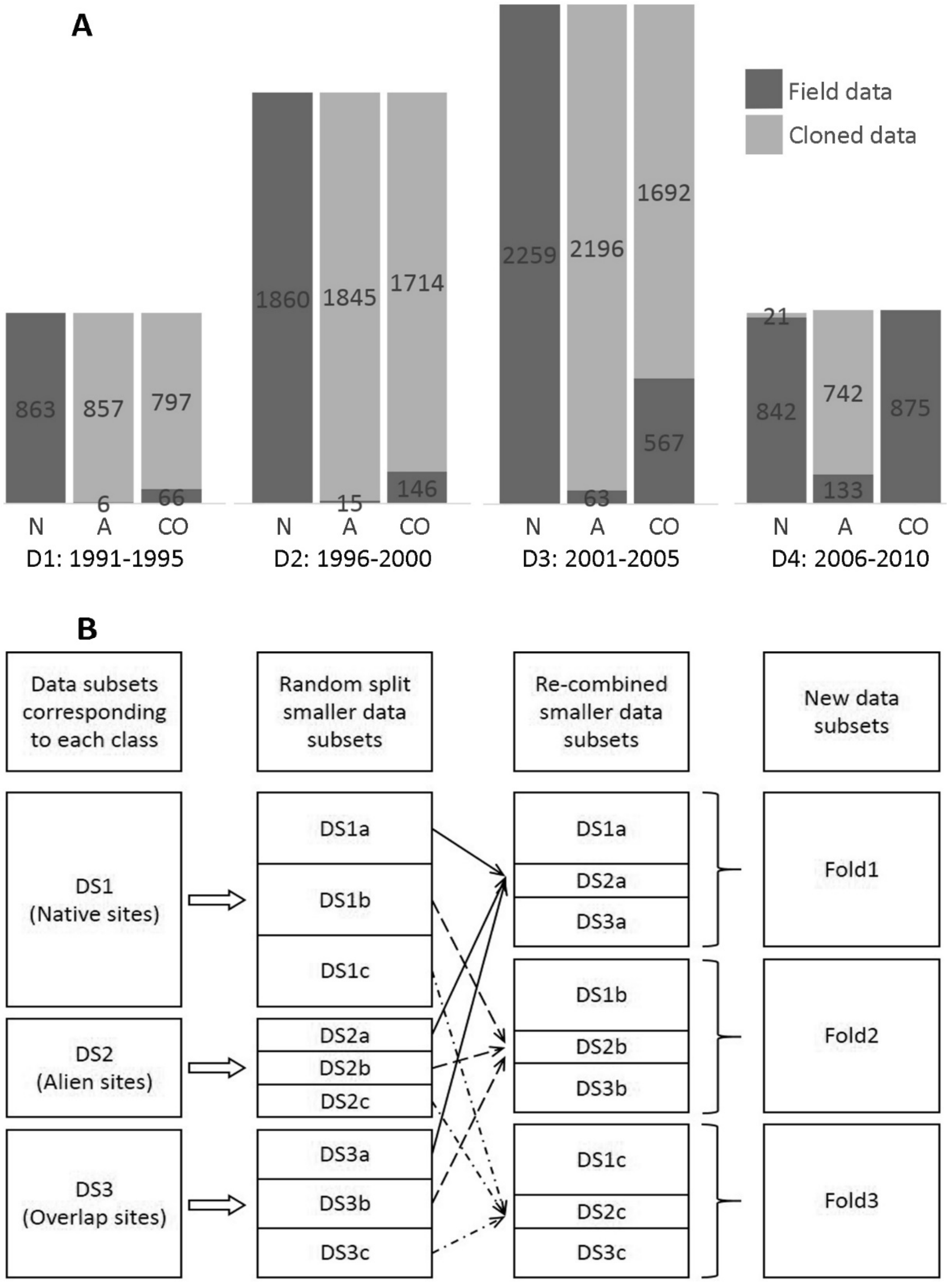


Fig. 2.3 A) Bar charts showing the number of the field and cloned data, which were combined together and used in the models. B) Schematic diagram illustrating the splitting procedure of stratified resampling approach.

into 3 smaller data subsets (i.e. DS1a, DS1b, DS1c; DS2a, DS2b, DS2c; DS3a, DS3b, DS3c). Thus, each of the smaller data subsets of the same class had exactly the same number of samples (n, e.g. nDS1a = nDS1b = nDS1c) or 1 sample more or less than the others (e.g. nDS1a ± 1). Finally, I recombined each smaller data subset to obtain new three equal-stratified data subsets, i.e. Fold1 (nDS1a + nDS2a + nDS3a), Fold2 (nDS1b + nDS2b + nDS3b) and Fold3 (nDS1c + nDS2c + nDS3c), which were later used to calibrate and validate the models. The schematic diagram showing this resampling procedure is depicted in Fig. 2.3B. To be consistent, I also made three replicates of the 3 fold CV for this resampling approach, by reshuffling the original data and then following the same procedure. The calibration and validation of the models was performed separately for each resampling approach.

For each scenario (hindcasting and forecasting), nine models (3 models of 3 fold CV × 3 replicates) were built for each resampling approach in each period. In this way, a performance value of the nine different models was calculated. From the field data, I built 36 models ([3 models of 3 fold CV hindcasting × 3 replicates × 2 resampling types] + [3 models of 3 fold CV forecasting × 3 replicates × 2 sampling types]) for each period. In the same way, 36 models were built based on the combined data for each period. Therefore, I totally built 144 models over the four periods for each data type (field and combined data). Kappa and CCI were used to evaluate the model performance.

3. Main Results

3.1 Macroinvertebrate communities and diversity patterns in the LMB

3.1.1 Overall macroinvertebrate communities

In total, 299 taxa, 196 genera, 90 families, 23 orders and 5 clades, accounting for a total of 21,810 individuals (inds), were identified; of which, 131 taxa (44%) and 6,481 inds (30%) were insects, 98 taxa (33%) and 10,603 inds (49%) were molluscs, 38 taxa (13%) and 2,054 inds (9%) were crustaceans, and 32 (10%) and 2,672 inds (12%) were annelids. The most commonly found insect orders were Diptera (37 taxa), Ephemeroptera (32), Odonata (22) and Trichoptera (20). For molluscs, most taxa belonged to the order Unionida (18), Veneroida (15) and Caenogastropoda (50); whereas for crustaceans, most taxa belonged to the order Decapoda (18) and Amphipoda (9). Annelids were mainly represented by the order Haplotaxida (15).

At the family level, insects were mainly represented by Chironomidae (Diptera; 24 taxa) and Gomphidae (Odonata; 14 taxa). Molluscs were characterized by Unionidae (Bivalvia; 18 taxa), Corbiculidae (Bivalvia; 14 taxa), Viviparidae (Gastropoda; 12 taxa) and Stenothyridae (Gastropoda; 9 taxa). Crustaceans and annelids were represented by Palaemonidae (10 taxa) and Naididae (15 taxa), respectively. Regardless the taxonomic groups, 36 macroinvertebrate families were represented by only one species (see Appendix T2).

Over all of 299 species reported, 20 were common (present in ≥16 samples or ≥25% occurrence), 106 were uncommon (present between 4-15 samples or between 6-24% occurrence) and 173 were rare (present in ≤3 samples or ≤5% occurrence). Alien species were not accessed due to the lack of distribution data and the limited taxonomical knowledge of each species. The most widely distributed species belonged to two insects: *Ablabesmyia* sp. (73% occurrence) and *Polypedilum* sp. (70%) and one was a mollusc, *Corbicula tenuis* (67%). In addition to being widely distributed, these 3 taxa were among the top 10 most abundant. Of the total individuals, *Ablabesmyia* sp. accounted for 2.9%, *Polypedilum* sp. for 3.8%, whereas the 3 most abundant species, *Corbicula leviuscula*, *Limnoperna siamensis* and *Corbicula tenuis*, accounted for 8.4%, 6.1% and 5.8%, respectively. The information of each species occurrence is provided in the Appendix T2.

3.1.2 Spatial community patterns and their relationship with environmental factors

3.1.2.1 Community clusters and indicator taxa

Based on the dissimilarity of macroinvertebrate abundance data and cluster analysis, the 63 sampling sites could be distinguished into four clusters (Fig. 3.1). Cluster I was situated in the Mekong delta in Vietnam; cluster IIa, along the upstream sites in Laos and Thailand; cluster IIb1, mostly in middle part in Cambodia and a few sites in Laos, Thailand and Vietnam; and cluster IIb2, mostly located in the tributaries of the LMB (Fig. 3.1A).

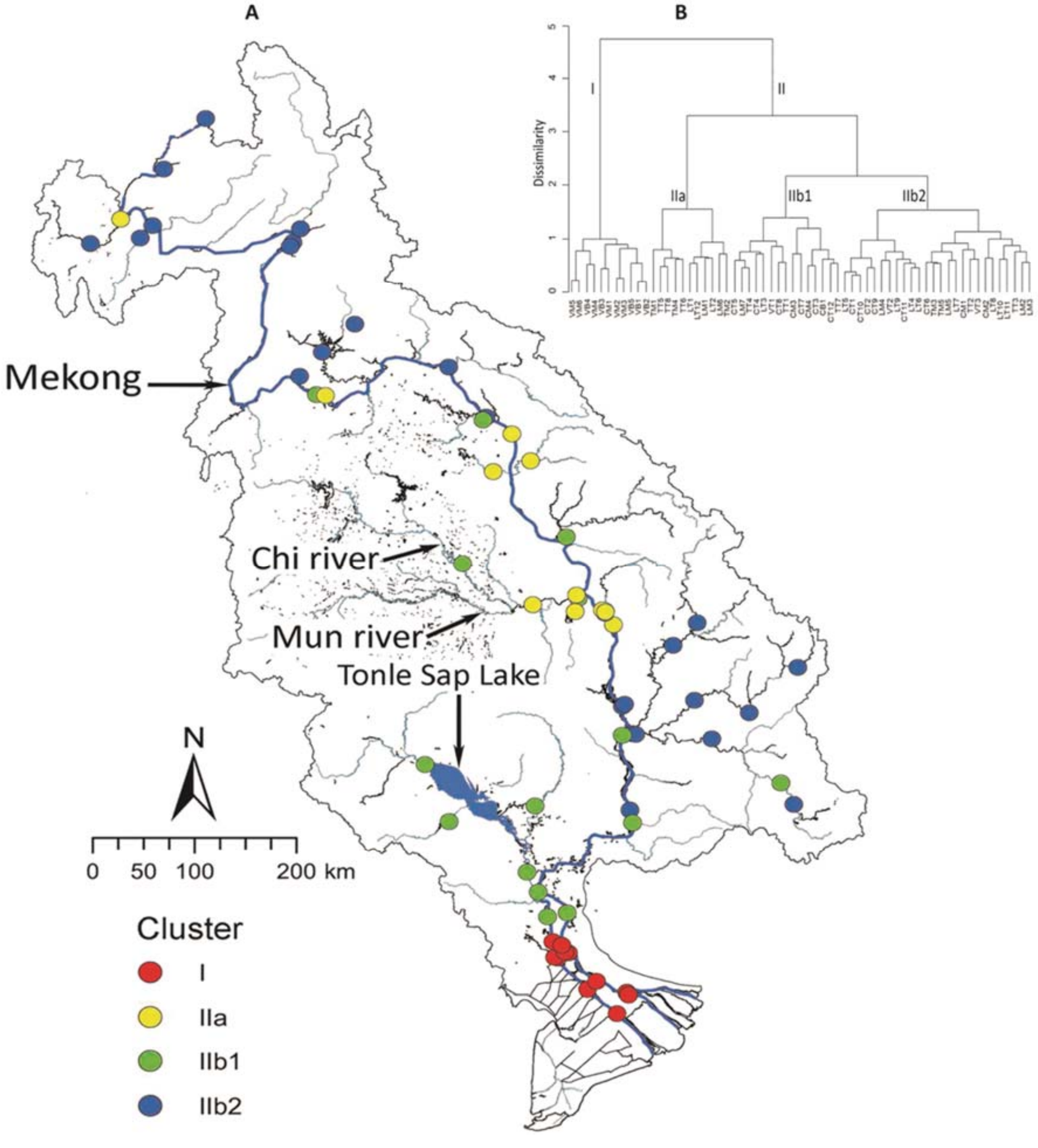


Fig. 3.1 Sampling sites and the four clusters, representing four macroinvertebrate assemblages, based on the cluster analysis (A), and the dendrogram showing sites belonging to the four clusters based on the dissimilarity and Ward's hierarchical clustering method (B).

The clusters represent four different macroinvertebrate assemblages, which were characterized by different environmental conditions, macroinvertebrate richness, abundance and diversity ($p<0.01$) (Fig. 3.2, Table 3.1). An increasing richness, abundance and diversity of macroinvertebrates was found from the upstream (i.e. tributaries) to the downstream assemblage (the delta). The richness and abundance of molluscs, crustaceans and annelids generally increased from the up to the downstream assemblage, while insects dominated the upstream and tributary assemblages.

Table 3.1 Mean value (and standard deviation) for environmental variables, richness, abundance and Shannon diversity of macroinvertebrate assemblage in each cluster.

	Cluster [n]			
Variable (unit)	I [11]	IIa [11]	IIb1 [15]	IIb2 [26]
Assemblage composition				
Richness*	54 (13)$^{-IIa, -IIb1,2}$	23 (11)	18 (7)	16 (9)
Abundance*	955 (526)$^{-IIa, -IIb1,2}$	251 (184)	233 (204)	193 (286)
Diversity (*H’)**	2.7 (0.7)$^{-IIa, -IIb1,2}$	2.1 (0.5)	2.0 (0.5)	1.9 (0.4)
Physical-chemical				
Altitude (m)*	6.6 (1.8)$^{-IIa, -IIb2}$	136 (77)	63 (67)$^{-IIb2}$	207 (1,539)
River width (m)*	1,057 (468)$^{-IIa, -IIb1,2}$	413 (372)	349 (412)	339 (375)
River depth (m)*	11.5 (3.5)$^{-IIa, -IIb1,2}$	5.0 (3.3)$^{-IIb2}$	5.0 (4.3)$^{-IIb2}$	2.5 (1.7)
Secchi depth (m)*	0.6 (0.2)$^{-IIb2}$	0.8 (0.4)	0.7 (0.5)	1.0 (0.7)
WT (°C)*	29.6 (0.5)$^{-IIa, -IIb2}$	26.0 (2.1)$^{-IIb1}$	28.9 (1.6)$^{-IIb2}$	26.4 (3.7)
DO (mg/L)*	6.2 (1.2)$^{-IIa, -IIb2}$	7.9 (0.5)$^{-IIb1}$	6.1 (1.7)$^{-IIb2}$	7.7 (0.7)
pH	7.8 (0.4)	7.6 (0.6)	7.5 (0.3)	7.6 (0.5)
EC (mS/m)*	17.8 (1.4)$^{-IIa, -IIb2}$	22.8 (6.0)$^{-IIb1}$	13.5 (6.4)$^{-IIb2}$	14.2 (9.8)
SAW (km^2)*	764,797 (4,714)$^{-IIa, -IIb1,2}$	180,454 (202,943)	187,351 (276,932)	123,341 (196,952)
Land cover (%)				
Agricultural land*	25.77 (0.29)$^{-IIb2}$	24.56 (26.7)	28.14 (24.82)$^{-IIb2}$	11.68 (12.71)
Bamboo*	0.47 (0)	0.17 (0.22)$^{-IIb2}$	0.61 (1.37)	2.14 (3.6)
Crops	5.48 (0.03)	5.53 (3.71)	3.97 (3.09)$^{-IIb2}$	8.59 (7.62)
Deciduous forests	10.02 (0.13)	15.1 (26.54)	20.55 (18.28)$^{-IIb2}$	9.01 (15.44)
Evergreen forests*	14.07 (0.06)	10.03 (5.5)$^{-IIb2}$	14.07 (10.52)	20.35 (15.62)
Glacier	0.1 (0)	0.11 (0.14)	0.04 (0.08)	0.08 (0.14)
Grassland	11.6 (0.06)	11.55 (13.89)	5.96 (7.32)	10.76 (13.22)
Inundated*	0.39 (0.01)$^{-IIa, -IIb2}$	0 $^{-IIb1}$	0.5 (0.96)$^{-IIb2}$	0
Mix_evg.dec	8.99 (0.05)	12.82 (11.64)	9.68 (6.49)	8.79 (7.9)
Plantations*	0.17 (0)	0.03 (0.07)	0.2 (0.35)	0.17 (0.3)
Regrowth*	0.88 (0.01)	0.31 (0.3)$^{-IIb2}$	0.98 (1.1)	1.25 (1.58)
Rocks*	0.6 (0)	1.46 (3.34)$^{-IIb1,2}$	0.23 (0.31)	0.4 (0.54)
Urban areas*	0.08 (0)	0.51 (1.49)$^{-IIb1,2}$	0.07 (0.06)	0.07 (0.11)
Water surface*	1.18 (0.01)$^{-IIb2}$	0.82 (1.57)	1.18 (1.26) $^{-IIb2}$	0.42 (0.87)
Wetland*	0.07 (0.01)	0.02 (0.02)	0.09 (0.18) $^{-IIb2}$	0.01 (0.02)
Wood- & shrub-land*	17.23 (0.07)	14.05 (10.68)$^{-IIb2}$	12.5 (7.91) $^{-IIb2}$	24.38 (15.76)

WT: water temperature, DO: dissolved oxygen, EC: electrical conductivity, SAW: the surface area of watersheds, Mix_evg.dec: mixed evergreen and deciduous forests. The number of samples [n] in each cluster is indicated between square brackets. * indicates ANOVA and Kruskal-Wallis Test for significant differences among clusters at $p<0.05$. Superscripts (IIa, IIb1, IIb2) indicate significant pair-wise comparisons between the corresponding cluster (each column) and superscript-labeled clusters (i.e. IIa, IIb1, IIb2) at $p<0.05$.

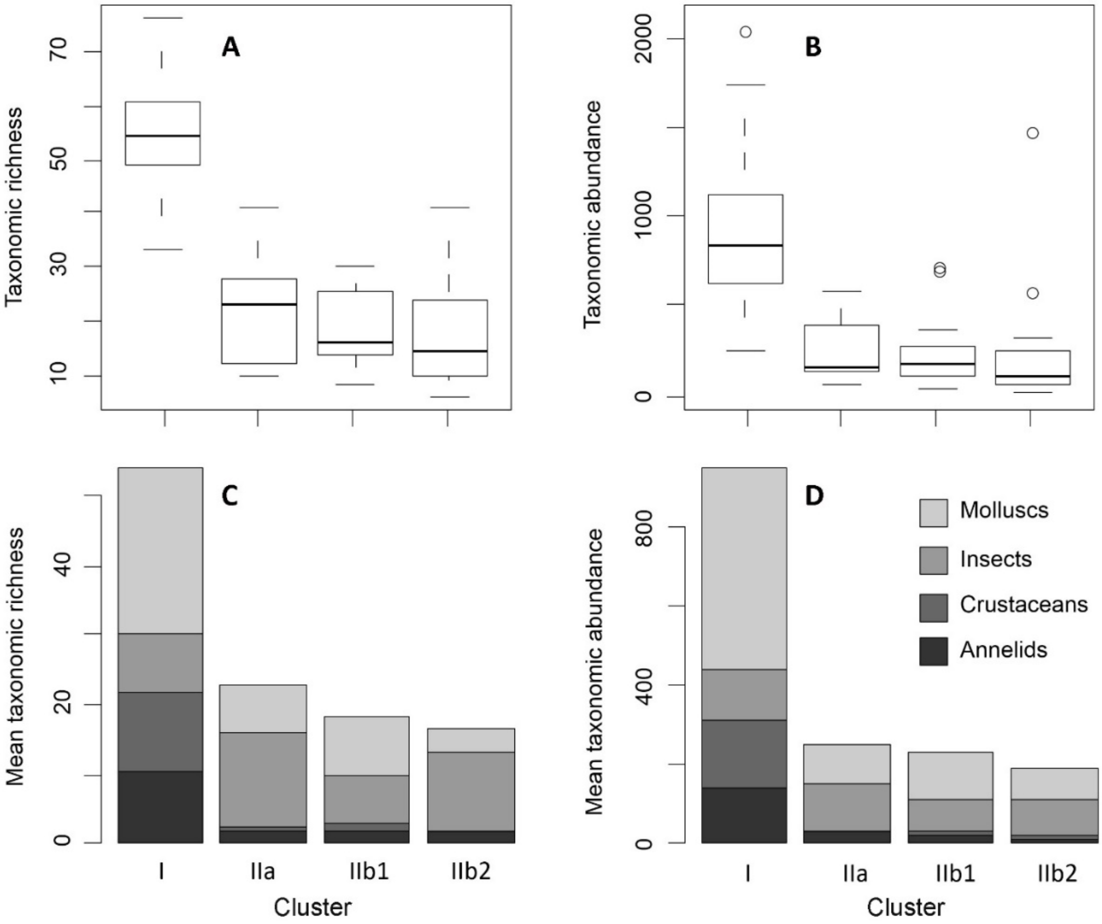


Fig. 3.2 Box and whisker plots of richness (A) and abundance (B) of macroinvertebrate assemblage in each cluster and its proportion of mean richness (C) and abundance (D) consisting of different components of macroinvertebrates.

The number of indicator species followed the overall trend of macroinvertebrate richness and abundance: the delta assemblage (I) were represented by 53 indicator species, most of which were molluscs, annelids and crustaceans. The upstream assemblage along the main channel were presented by 14 indicator taxa, most of which were insects. The in-between assemblage (IIb1, between the delta and main upstream assemblages) and the tributaries (IIb2) were represented by two different indicator taxa. The detailed information on indicator taxa for each assemblage is provided in Table 3.2.

Table 3.2 List of indicator taxa (and their indicator values, IndVal) of macroinvertebrate assemblage in each cluster.

Cluster I		
Annelid	IndVal	*p*-value
Aeolosoma bengalense	0.52	0.010
Aulodrilus prothecatus	0.67	0.005
Chaetogaster langi	0.85	0.005
Chaetogaster limnaei limnaei	0.6	0.005
Dero pectinata	0.67	0.005
Dero sp.	0.74	0.005
Dero sp.1	1	0.005
Dero sp.2	0.95	0.005
Lumbriculidae sp.	0.6	0.005
Namalycastis longicirris	0.9	0.005
Orbinia johnsoni	0.52	0.010
Polydora sp.	0.67	0.005
Crustacean		
Corophium minutum	0. 8	0.005
Corophium sp.	0.67	0.005
Cyathura carinata	0.74	0.005
Cyathura truncata	0.57	0.005
Decapoda sp.	0.91	0.005
Eohaustorius sp.	0.6	0.005
Eohaustorius tandeensis	0.67	0.005
Gammarus sp.	0.6	0.005
Grandidierella lignorum	0.78	0.005
Grandidierella vietnamica	1	0.005
Hyale hawaiensis	0.67	0.005
Hyale sp.	0.85	0.005
Kamaka sp.	0.6	0.005
Macrobrachium equidens	0.6	0.005
Melita sp.	0.82	0.005
Monocorophium sp.	0.91	0.005
Palaemon curvirostris	0.6	0.005
Mollusc		
Afropisidium clarkeanum	0.73	0.005
Angulyagra polyzonata	0.6	0.005
Angulyagra sp.	0.9	0.005
Bithynia siamensis	0.67	0.005
Corbicula baudoni	0.87	0.005
Corbicula bocourti	0.74	0.005
Corbicula leviuscula	0.97	0.005
Corbicula moreletiana	0.86	0.005
Corbicula sp.	0.95	0.005
Gastropoda sp.	0.74	0.005
Hyriopsis bialatus	0.64	0.005
Limnoperna siamensis	0.99	0.005
Limnoperna sp.	0.95	0.005
Lymnaea viridis	0.94	0.005
Mekongia swainsoni swainsoni	0.67	0.005
Sinomytilus harmandi	0.9	0.005
Stenothyra annandalei	0.6	0.005
Stenothyra glabrata	0.85	0.005
Trochotaia trochoides	0.52	0.005
Cluster I (continued)		
Insect	IndVal	*p*-value
Arigomphus sp.	0.67	0.005
Cricotopus sp.	1	0.005
Clinotanypus sp.	0.52	0.005
Nectopsyche sp.	0.67	0.005
Sigara sp.	0.6	0.005

Cluster IIa		
Annelid	IndVal	*p*-value
Oligochaeta sp.	0.99	0.005
Polychaeta sp.1	0.6	0.005
Mollusc		
Corbicula sp.	0.88	0.005
Hubendickia sp.	0.6	0.010
Kareliania sp.	0.52	0.010
Scaphula sp.	0.52	0.010
Stenothyra sp.	0.6	0.005
Insect		
Anagenesia sp.	0.67	0.005
Caenoculis sp.	0.52	0.010
Caenodes sp.	0.74	0.005
Choropterpes sp.	0.51	0.005
Dipseudopsis sp.	0.69	0.005
Heterocloeon sp.	0.52	0.010
Micronecta sp.	0.6	0.005

Cluster IIb1		
Mollusc	IndVal	*p*-value
Filopaludina filopaludina filosa	0.45	0.025
Insect		
Pentagenia sp.	0.62	0.010

Cluster IIb2		
Annelid	IndVal	*p*-value
Naididae sp.	0.76	0.005
Insect		
Gomphidae sp.	0.56	0.010

3.1.2.2 Relationship between community clusters and environmental factors

The results of the LDA model used to discriminate the macroinvertebrate assemblages based on the physical-chemical variables and land cover types are shown in Fig. 3.3. Along axis 1, assemblage I was situated opposite to assemblage IIa and IIb2. Assemblage I was positively correlated with the surface area of watershed, river depth, river width and water temperature, but negatively associated with altitude and dissolved oxygen. Whereas assemblage IIa was positively correlated with electrical conductivity and urban areas, and assemblage IIb2 was positively linked to altitude, DO, Secchi depth, wood-/shrub-land and evergreen forests. Based on axis 1 and 3, assemblage IIb1 was positively linked to inundated, wetland and agricultural areas (Fig. 3.3).

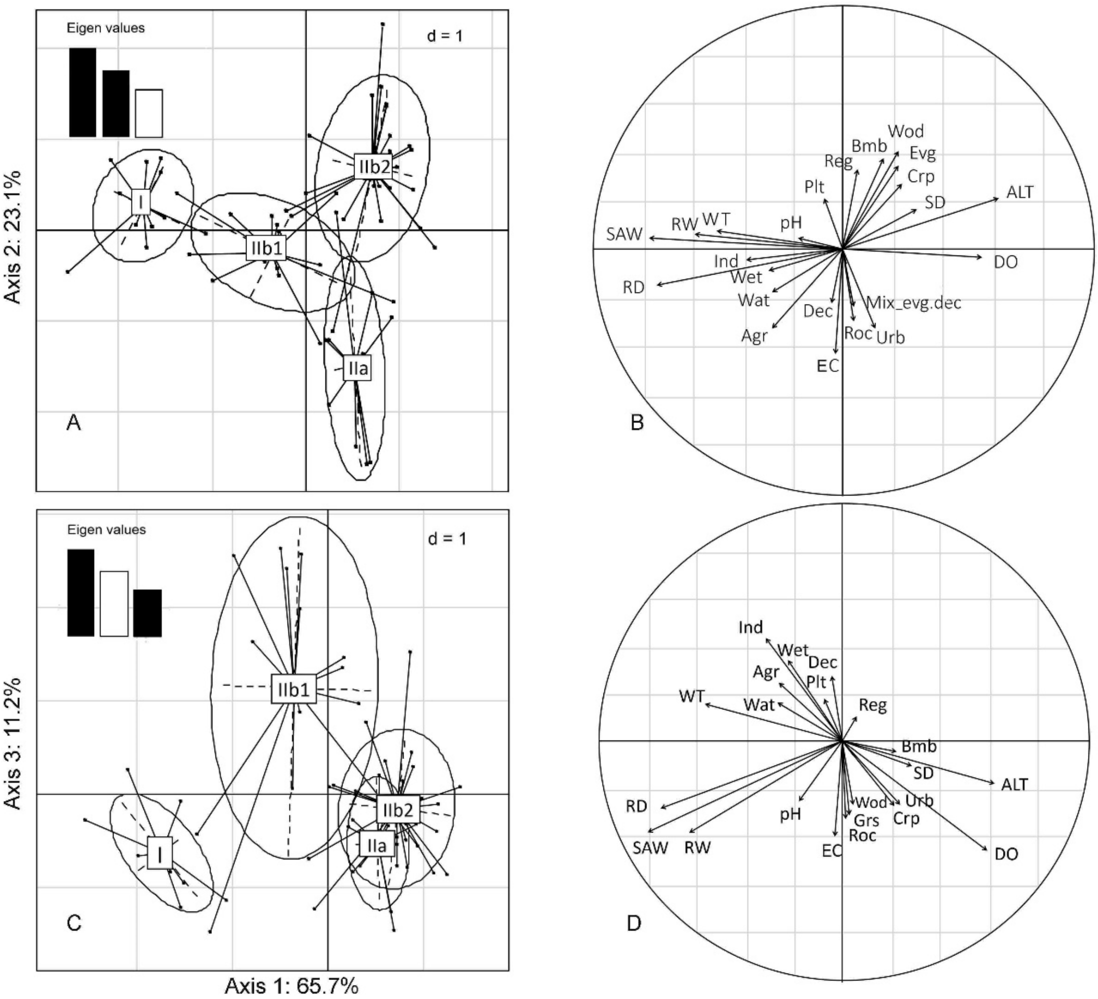


Fig. 3.3 Results from the LDA discriminating the four clusters (I, II, IIb1, IIb2), representing four macroinvertebrate assemblages, using Axes 1, 2 and 3 that explained the indicated percentage of the total variance in the data (A, C), and correlations of the environmental factors to the corresponding axes (B, D). ALT: altitude, RW: river width, RD: river depth, SD: Secchi depth, WT: water temperature, DO: dissolved oxygen, EC: electrical conductivity, SAW: the surface area of watersheds, Agr: agricultural land, Bmb: bamboos, Crp: crops, Dec: deciduous forests, Evg: evergreen forests, Grs: grassland, Ind: inundated, Mix_evg.dec: mixed evergreen and deciduous forests, Plt: plantations, Reg: regrowth, Roc: rocks, Urb: urban areas, Wat: water, Wet: wetland, Wod: wood- & shrub-land.

3.1.3 Macroinvertebrate diversity and its relation to environmental factors

3.1.3.1 Diversity and variation in important taxa

The diversity of macroinvertebrates remained relatively high across the LMB. The mean richness and abundance at each site was 23 species (range: 6-74 species) and 346 inds (range: 13-2009 inds). Alpha (α) diversity (*H')* at each site was 2.1 (range: 0.8-3.3). Beta diversity of the global macroinvertebrate communities (i.e. the communities that include all component communities: molluscs, crustaceans, annelids and insects), measured as the total variance, was exceptionally high, at BD_{Total} =0.80 on a 0-to-1 scale. When β diversity of each component community was computed separately, the total variance of mollusc communities was the highest (BD_{MTotal} =0.78), followed by insect (BD_{ITotal} =0.74) and annelid communities (BD_{ATotal} =0.72). Crustacean communities had the lowest total variation (BD_{CTotal} =0.38).

A total of 60 macroinvertebrates were identified as the important species (i.e. the species that had SCBD indices larger than the mean SCBD (0.003). Among them, 29 species belonged to insects, 18 to molluscs, 7 to annelids and 6 to crustaceans (Table 3.3). The SCBD values are small because they are relative to the total sum of squares in the community composition table and sum to 1. High SCBD indices indicated taxa that have high variance across sites, and thus greatly contributed to the BD_{Total}. The 3 highest SCBD indices belonged to insect taxa: *Polypedilum* sp. (I033, SCBD =0.054), *Ablabesmyia* sp. (I017, 0.039), *Cryptochironomus* sp. (I024, 0.037), followed by *Corbicula tenuis* (mollusc, B37; 0.037), *Goeldichironomus* sp. (insect, I028; 0.035) and *Corbicula leviuscula* (mollusc, B34; 0.034).

LCBD indices of the global communities and of component communities, which indicate the uniqueness in taxonomic composition at the sites, are provided in Appendix T3. The LCBD values are scaled to add up to 1 over the whole study; the larger the number of sample sites included, the smaller the LCBD value yielded. Large or small LCBD values indicate the sites that respectively contribute more or less than the mean to β diversity.

3.1.3.2 Relationship between diversity and environmental factors

Based on the AIC criterions and stepwise selection of the multiple regression models, only river depth, surface area of watersheds, electrical conductivity, and Secchi depth, among all measured environmental variables, remained significantly associated with the global LCBD indices. River depth, surface area of watersheds and electrical conductivity showed a positive association, while Secchi depth had a negative association. The four variables accounted for 29% (adjusted R^2 = 0.29) of the variation of global LCBD indices. Among the component communities, only LCBD_I and LCBD_M indices that best explained the variation of the global LCBD indices (adjusted $R^2 = 0.84$).

Table 3. 3 List of taxa with high SCBD indices (above the overall mean). The first letter of each taxon code represents the component of macroinvertebrate communities (A: annelid, C: crustacean, B and G: molluscs, and I: insect).

Code	Species name	SCBD	Code	Species name	SCBD
A31	Naididae sp.	0.0290	I033	*Polypedilum* sp.	0.0540
A17	*Limnodrilus hoffmeisteri*	0.0280	I017	*Ablabesmyia* sp.	0.0390
A11	Oligochaeta sp.	0.0190	I024	*Cryptochironomus* sp.	0.0373
A16	*Branchiura sowerbyi*	0.0190	I028	*Goeldichironomus* sp.	0.0350
A04	*Dero* sp.1	0.0090	I020	*Chironomus* sp.	0.0320
A13	*Chaetogaster* sp.	0.0060	I079	*Pentagenia* sp.	0.0250
C03	*Grandidierella vietnamica*	0.0080	I051	*Anagenesia* sp.	0.0180
C05	*Melita* sp.	0.0080	I014	*Culicoides* sp.	0.0170
C01	*Corophium* sp.	0.0070	I146	Philopotamidae sp.	0.0170
C04	*Kamaka* sp.	0.0050	I059	*Caenis* sp.	0.0140
C20	*Cyathura truncata*	0.0050	I026	*Einfeldia* sp.	0.0110
C18	*Macrobrachium* sp.	0.0034	I078	Palingeniidae sp.	0.0100
B37	*Corbicula tenuis*	0.0371	I013	*Bezzia* sp.	0.0090
B34	*Corbicula leviuscula*	0.0340	I061	*Caenodes* sp.	0.0090
B03	*Limnoperna siamensis*	0.0290	I133	*Macronema* sp.	0.0080
B36	*Corbicula* sp.	0.0270	I018	Chironomidae sp.	0.0070
B32	*Corbicula lamarckiana*	0.0230	I035	*Pseudochironomus* sp.	0.0070
G22	*Stenothyra koratensis holosculpta*	0.0230	I037	*Sergentia* sp.	0.0070
B20	*Uniandra contradens ascia*	0.0160	I066	*Ephemera* sp.	0.0070
G25	*Stenothyra mcmulleni*	0.0150	I069	*Cladopelma* sp.	0.0070
G38	*Hubendickia crooki*	0.0150	I038	*Smittia* sp.	0.0060
B26	*Corbicula blandiana*	0.0110	I023	*Cricotopus* sp.	0.0050
G39	*Hubendickia* sp.	0.0100	I086	*Diplonychus rusticus*	0.0050
B04	*Limnoperna* sp.	0.0090	I088	*Microtendipes* sp.	0.0050
B38	*Sinomytilus harmandi*	0.0070	I153	*Naucoris* sp.	0.0050
G51	*Bithynia* sp.	0.0060	I119	*Progomphus* sp.	0.0050
G54	*Kareliania* sp.	0.0060	I127	*Dipseudopsis* sp.	0.0050
G12	*Mekongia swainsoni flavida*	0.0050	I065	*Eatonigenia* sp.	0.0040
G43	*Pachydrobia* sp.	0.0040	I112	Gomphidae sp.	0.0040
B29	*Corbicula cyreniformis*	0.0034	I150	Psychomyiidae sp.	0.0040

3.2 Macroinvertebrate communities and diversity patterns in Flemish rivers

3.2.1 Overall macroinvertebrate communities

During the 20-year monitoring (1991-2010) across fluvial systems in Flanders, 207 taxa (6,192,056 inds) belonging to 145 families, 25 orders and 4 clades were identified. Only 123 taxa were identified to genus level; the reaming 84 taxa were identified to only at the family level. Among all recorded taxa, 131 taxa (63%) and 1,752,885 inds (28%) belonged to insects, 34 taxa (16%) and 628,468 inds (10%) to molluscs, 21 taxa (10%) and 1,589,809 inds (26%) to crustaceans, and 21 taxa (10%) and 2,220,894 inds (36%) to annelids. The dominant insect orders were Diptera (25 taxa), Odonata (25), Hemiptera (22), Trichoptera (18) and

Ephemeroptera (17). Molluscs were mainly represented by the clade Hygrophila (17 taxa), Caenogastropoda (6) and the order Veneroida (5) and Unionida (4). Crustaceans and annelids were dominated by the order Decapoda (6 taxa) and Amphipoda (4), and Rhynchobdellida (7) and Arhynchobdellida (5), respectively.

The most occurring families of insects were Corixidae (Hemiptera, 9 taxa), Coenagrionidae (Odonata, 8), of molluscs was Planorbidae (Hygrophila, 11) and of annelids was Glossiphoniidae (Arhynchobdellida, 5). Crustaceans were identified only to family level, and that did not reveal which family dominated the communities. In total, 73 alien taxa have been collected across river system in Flanders (see Boets et al. 2016).

3.2.2 Spatio-temporal community composition and environmental factors

3.2.2.1 Spatio-temporal community composition

From the past to the recent period, there was a gradual increase in macroinvertebrate richness and abundance (Fig. 3.4, Table 3.4). The average richness and abundance per site for the four periods (D1, D2, D3 and D4) were: 14±0.1 taxa and 380±32 inds, 14±0.2 taxa and 279±20 inds, 16+0.2 taxa and 456±24 inds, and 16±0.2 taxa and 1009±70 inds, respectively. The descriptive information on the richness and abundance of the global macroinvertebrate communities and of each component were provided in Table 3.4.

Table 3.4 Mean and standard error (SE) of richness and abundance of global and each component community for each period.

Communities	Richness (SE)				Abundance (SE)			
	D1	D2	D3	D4	D1	D2	D3	D4
Global	13.6 (0.12)	14.4 (0.15)	16.4 (0.15)	15.7 (0.19)	380 (32)	279 (20)	456 (24)	1009 (70)
Annelids	3.1 (0.04)	3.3 (0.03)	3.6 (0.03)	3.4 (0.04)	204 (28)	125 (16)	108 (10)	238 (17)
Crustaceans	1.6 (0.02)	1.5 (0.02)	1.7 (0.02)	1.7 (0.03)	55 (9)	46 (6)	138 (14)	311 (64)
Insects	6.2 (0.11)	6.7 (0.09)	7.8 (0.1)	6.8 (0.1)	83 (8)	79 (5)	150 (11)	335 (15)
Molluscs	2.8 (0.07)	2.9 (0.06)	3.4 (0.05)	3.8 (0.07)	38 (8)	29 (2)	60 (6)	125 (12)

In the first two periods in the 1990s, significantly seasonal differences (i.e. Spring vs Summer vs Autumn) was observed for macroinvertebrate abundance (Friedman chi-squared F =14.8; p<0.001) in the early 1990s (D1), and for macroinvertebrate richness (F =40.6, p<0.001) and abundance (F =20.3, p<0.015) in the late 1990s (D2). During the 2000s, no significant difference was found for macroinvertebrate richness nor abundance. The richness and abundance between each season across the four periods, e.g. Spring (D1) vs Spring (D2) vs Spring (D3) vs Spring (D4), also showed significant difference at p<0.01 (Fig. 3.4).

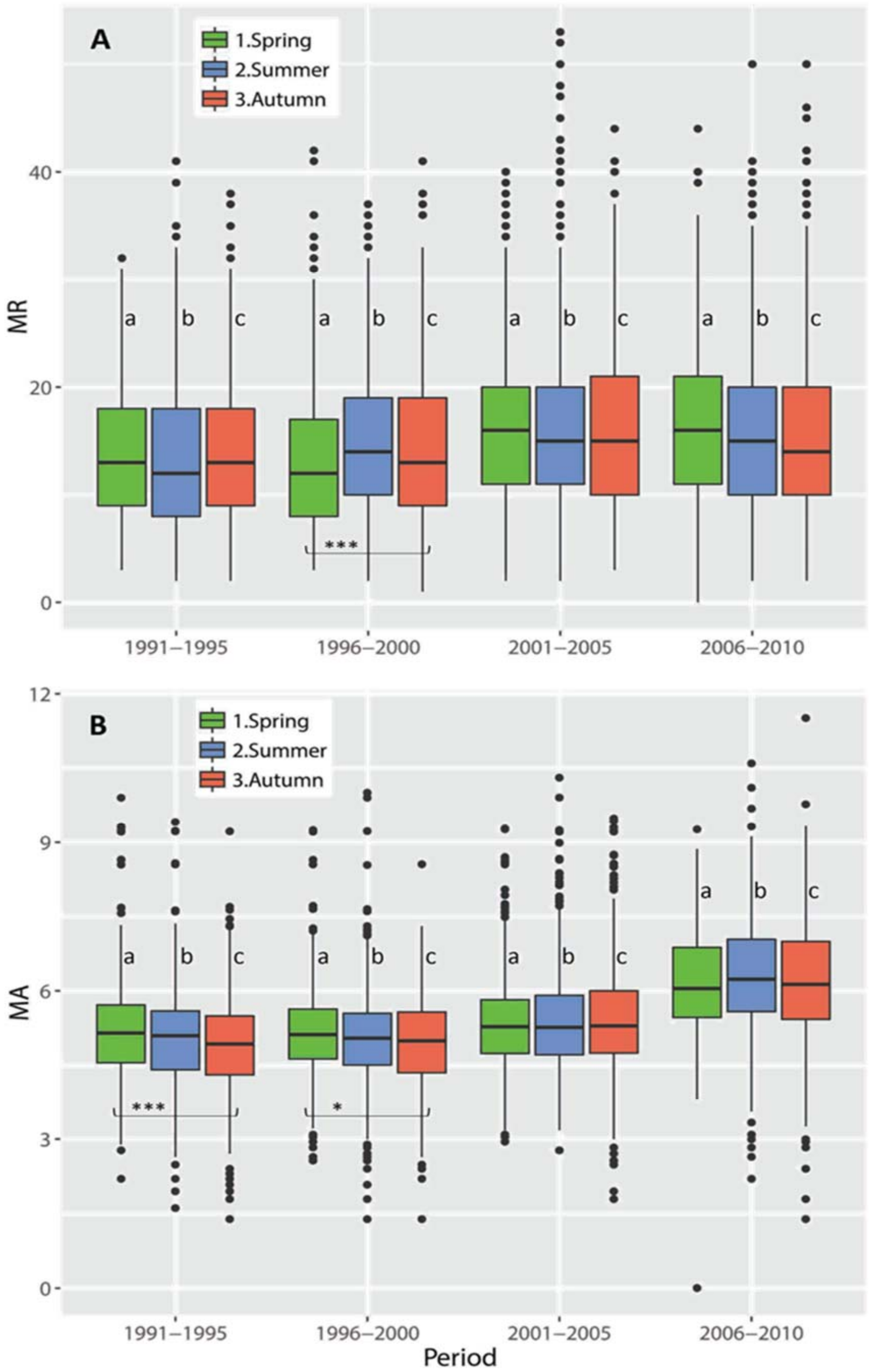


Fig. 3.4 Box and whisker plots of macroinvertebrate richness (MR, A) and log+1 abundance (MA, B) between seasons in each period. Significant difference in richness and abundance for each season across the four periods is indicated as “a”, “b”, “c”. * p<0.05, *** p<0.001.

3.2.2.2 Relationship between community composition and environmental factors

Based on the RDA analysis, the environmental variables explained a small amount of variance (Adjusted $R^2 = 0.09$) of the two-decade community data. Most taxa were more related to sites having high values of DO and low values of other water quality variables. Exceptions were found for four taxa which were negatively and strongly associated with DO concentration. An insect taxon (Chironomidae *thummi-plumosus*, I017) were positively associated with sites having high values of ammonium, total phosphate, phosphate, COD, and a crustacean (Asellidae, C15) and two annelids (Naididae, A11; Helobdella, A15) were linked to sites having high values of ammonium and nitrite. Over time, macroinvertebrate samples collected during the early 1990s (D1) were positively related to concentration of ammonium, total phosphate, orthophosphate, COD, but negatively linked to DO concentration. In the late 1990s (D2), samples are more related to high values of sinuosity, and in the early (D3) and late (D4) 2000s, samples were more related to high values of DO and pH, respectively (Fig. 3.5).

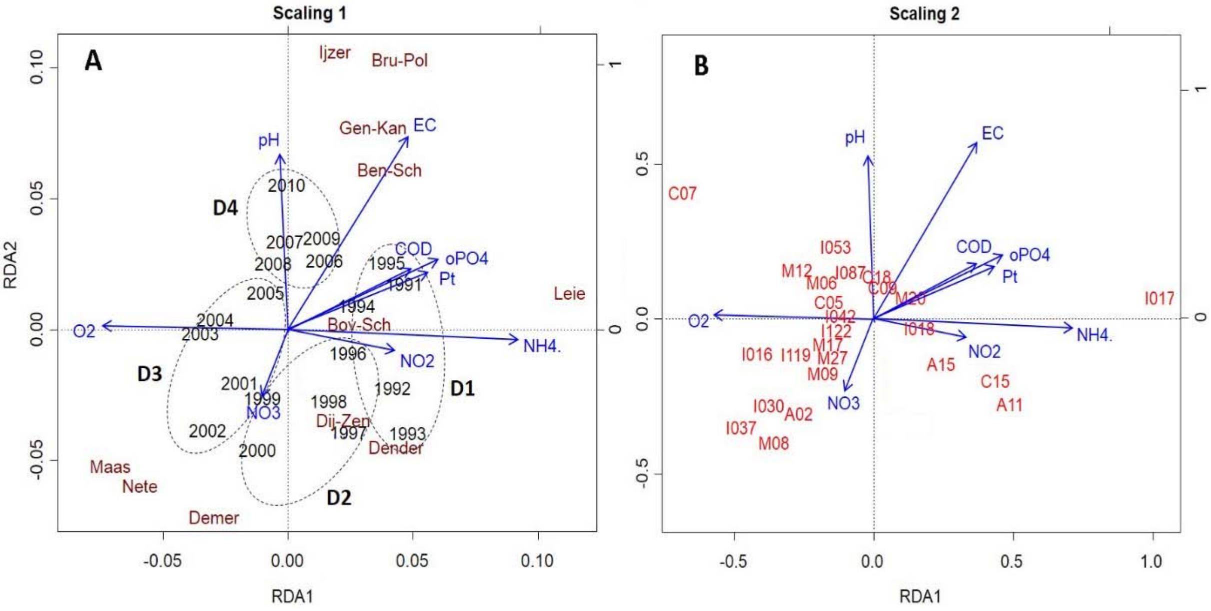


Fig. 3.5 RDA ordination plots showing the association of macroinvertebrates with environmental variables. A) Plot showing sampling years and river basins; Name of river basins: Ijzer, Leie, Maas, Nete, Demer, Dender, Bru-Pol: Brugse Polders, Gen-Kan: Gentse Kanalen, Ben-Sch: Beneden-Schelde, Bov-Sch: Boven-Schelde, and Dij-Zen: Dijle Zenne. B) Plot showing macroinvertebrate taxa and environmental factors for the same analysis. Code of taxa with small loadings score were removed to improve legibility; see Table 3.5 for full names of the taxa.

3.2.3 Spatio-temporal diversity pattern and its relation to environmental factors

3.2.3.1 Diversity and variation in important taxa

Mean α diversity (Shannon-Weiner, *H'*) of the global communities gradually increased from the early 1990s (*H'*=0.6±0.02) to the early 2000s (*H'*=1.9±0.01), but decreased in the late 2000s (*H'*=1.6±0.02) (Fig. 3.6A). For each component community, α diversity of insects was the highest, followed by molluscs' and annelids', while crustaceans had the lowest diversity. However, the trend of α diversity of each component followed the global diversity trend.

Based on the Friedman test, α diversity of the global communities among the four periods was significantly different (Friedman chi-squared F =308, p<0.001). Significant seasonal variation in α diversity (F >25.1, p<0.001) was also detected in each period of the 1990s (D1 and D2), but not in the 2000s (Fig. 3.6B). When each season was compared across the four periods, significant difference was always observed (Fig. 3.6B). The information on α diversity of the component community is summarized on Fig. 3.6A.

From the past (D1, 1991) to the most recent (D4, 2010) period, an increasing trend in β diversity (the total variation) of global macroinvertebrate communities was observed. The mean BD_{Total} for the corresponding period D1, D2, D3 and D4 was 0.52±0.07, 0.50±0.03, 0.54±0.02 and 0.59±0.01 (Fig. 3.7A). Based on seasons, the total variation of the global communities was lower in spring, compared to summer in autumn, for the first two periods. The BD_{Total} for the first two periods in spring, summer and autumn were: 0.44, 0.54 and 0.58 (for D1), and were 0.46, 0.53 and 0.51 (for D2), respectively. For the later two periods (2001-2010), the amount of seasonal total variation in each period was relatively similar (Fig. 3.7B).

When β diversity of each component community was separately computed for each period, molluscs always had the highest variation, followed by insects and crustaceans. Annelids always had the lowest variation. Over the four periods, β diversity of molluscs and insects increased from the past to the most recent period, whereas β diversity of annelids and crustaceans fluctuated. The information on β diversity of each component community in each period is shown in Table 3.5.

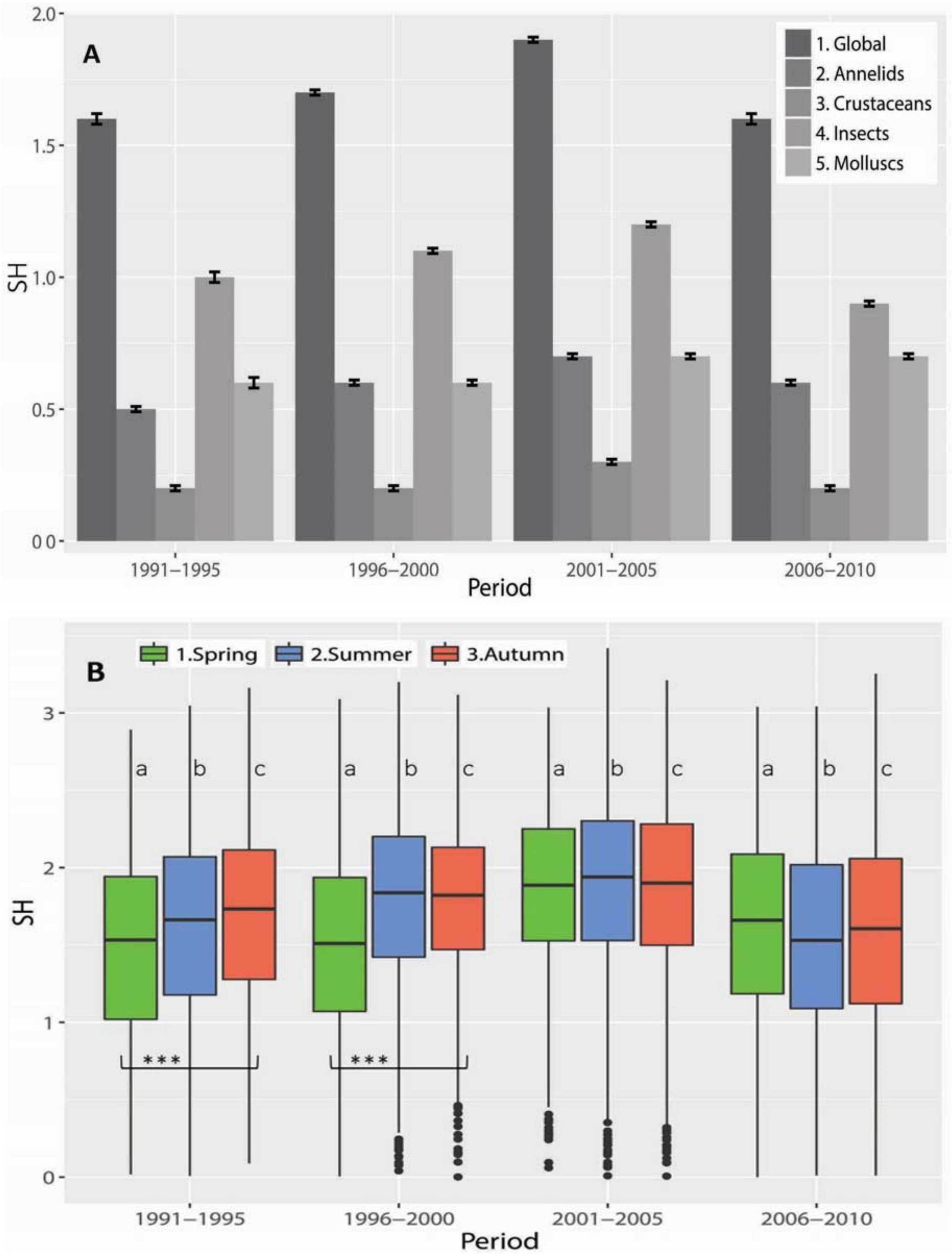


Fig. 3.6 Bar and standard error plots showing α diversity of the global and component communities in each period (A), and box and whisker plot showing α diversity (SH) of global communities between seasons (B). Significant difference in α diversity for each season across the four periods is shown as “a”, “b”, “c” (B). *** p<0.001.

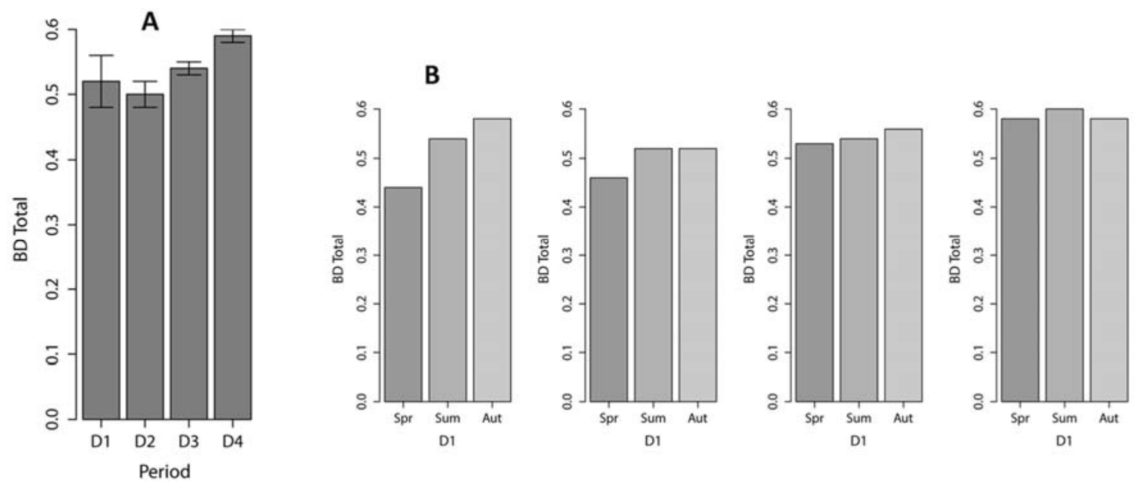


Fig. 3.7 Bar and standard error plots showing β diversity of global communities between periods (A) and between seasons of each period (B).

Table 3.5 Beta diversity of component communities between seasons in each period.

		Component communities			
Period	Season	Annelids	Crustaceans	Insects	Molluscs
D1: 1991-1995	Spring	0.19	0.39	0.47	0.57
	Summer	0.33	0.36	0.47	0.60
	Autumn	0.34	0.43	0.52	0.58
	Mean ± SE	*0.28±0.05*	*0.39±0.02*	*0.48±0.02*	*0.58±0.01*
D2: 1996-2000	Spring	0.20	0.37	0.46	0.60
	Summer	0.25	0.36	0.51	0.59
	Autumn	0.24	0.34	0.55	0.56
	Mean ± SE	*0.23±0.02*	*0.36±0.01*	*0.50±0.02*	*0.58±0.01*
D3: 2001-2005	Spring	0.23	0.38	0.52	0.63
	Summer	0.28	0.35	0.49	0.64
	Autumn	0.28	0.38	0.54	0.66
	Mean ± SE	*0.27±0.02*	*0.37±0.01*	*0.51±0.01*	*0.65±0.01*
D4: 2006-2010	Spring	0.31	0.44	0.51	0.66
	Summer	0.32	0.42	0.52	0.71
	Autumn	0.25	0.36	0.56	0.68
	Mean ± SE	*0.30±0.02*	*0.41±0.02*	*0.53±0.02*	*0.68±0.02*

SCBD indices identified 30 to 40 important taxa (i.e. the taxa having SCBD indices larger than the mean SCBD in each season) that contributed most to β diversity of the global communities in each period. Five taxa namely Naididae (annelid), Gammaridae and Asellidae (crustaceans), and the Chironomidae *thummi-plumosus* and Chironomidae non *thummi-plumosus* groups (insects) always highly contributed (i.e. ≥5% of variance) to global β diversity. Table 3.6 lists all the important taxa which were identified for each period.

Table 3.6 List of important taxa that the key contributors to the β diversity of global communities. The most important taxa that always had a high contribution were in bold. sp: spring, sm: summer, at: autumn.

		Period											
		D1: 1991-1995			D2: 1996-2000			D3:2001-2005			D4: 2006-2010		
Code	Taxon	sp	sm	at	sp	sm	at	sp	sm	at	sp	sm	at
A02	*Erpobdella*	0.029	0.021	0.015	0.034	0.023	0.032	0.024	0.028	0.027	0.014	0.018	0.019
A08	Enchytraeidae	0.031	0.054	0.040	0.009	-	-	-	-	-	-	-	-
A09	Haplotaxidae	0.016	-	-	-	-	-	-	-	-	-	-	-
A10	Lumbricidae	0.005	-	-	0.007	-	-	0.005	-	-	-	-	-
A11	**Naididae**	**0.145**	**0.130**	**0.110**	**0.142**	**0.098**	**0.093**	**0.093**	**0.077**	**0.070**	**0.093**	**0.101**	**0.103**
A12	Lumbriculidae	0.017	0.006	0.005	-	0.005	0.007	0.007	0.005	0.008	-	-	-
A13	*Glossiphonia*	0.012	0.011	0.011	0.015	0.015	0.016	0.016	0.020	0.018	0.011	0.016	0.011
A15	*Helobdella*	0.023	0.029	0.025	0.025	0.032	0.036	0.024	0.031	0.037	0.018	0.032	0.027
C04	Limnadiidae	-	-	-	-	-	-	-	-	-	0.024	-	-
C05	Corophiidae	0.012	0.005	0.007	0.012	0.007	-	0.009	0.005	-	0.019	0.019	0.007
C07	**Gammaridae**	**0.067**	**0.051**	**0.078**	**0.075**	**0.071**	**0.090**	**0.085**	**0.068**	**0.092**	**0.101**	**0.087**	**0.080**
C09	Palaemonidae	-	0.012	0.023	-	-	0.005	-	-	0.014	-	-	-
C15	**Asellidae**	**0.070**	**0.076**	**0.072**	**0.105**	**0.075**	**0.071**	**0.053**	**0.076**	**0.081**	**0.070**	**0.096**	**0.083**
C18	Mysidae	-	0.021	0.034	-	0.014	0.017	-	0.008	0.018	-	0.010	-
C21	*Ostracoda*	-	-	0.005	-	-	0.005	0.013	0.012	0.025	0.010	-	0.005
I001	Dryopidae	-	-	-	-	-	-	-	-	-	0.006	0.006	0.005
I002	Dytiscidae	0.020	0.013	0.014	0.007	0.013	0.011	0.012	0.010	0.008	-	-	-
I005	Haliplidae	-	-	0.005	-	0.005	0.005	-	0.007	0.006	-	-	-
I007	Hydrophilidae	-	-	-	-	0.006	-	0.006	0.010	-	-	-	-
I014	Ceratopogonidae	0.005	-	-	-	-	-	0.018	-	-	0.011	-	-
I016	**Chironomidae. Non *thummi-plumosus***	**0.076**	**0.075**	**0.065**	**0.084**	**0.065**	**0.064**	**0.064**	**0.060**	**0.050**	**0.076**	**0.074**	**0.070**
I017	**Chironomidae. *Thummi-plumosus***	**0.076**	**0.105**	**0.079**	**0.095**	**0.082**	**0.090**	**0.080**	**0.094**	**0.066**	**0.141**	**0.120**	**0.128**
I018	Culicidae	-	0.008	0.009	-	0.006	0.008	-	0.009	0.009	-	0.014	0.014
I025	Psychodidae	-	-	-	0.005	0.006	-	0.011	0.011	0.007	-	-	0.008
I030	Simuliidae	0.008	0.015	0.010	0.019	0.030	0.031	0.033	0.030	0.033	0.043	0.042	0.028
I035	Limoniidae	0.006	0.006	-	0.005	0.010	0.006	0.008	0.007	-	-	-	-
I037	Baetis	0.021	0.017	0.009	0.027	0.041	0.030	0.039	0.034	0.023	0.022	0.036	0.020
I039	*Cloeon*	0.005	0.020	0.032	0.007	0.023	0.022	0.010	0.017	0.021	0.013	0.008	0.019
I042	*Caenis*	0.011	-	-	0.008	-	-	0.008	0.007	-	0.013	0.005	-
I053	*Potamopyrgus*	0.010	0.018	0.014	0.019	0.017	0.019	0.019	0.018	0.031	0.022	0.023	0.030
I061	*Micronecta*	-	-	-	0.007	0.010	-	0.007	0.006	-	0.008	0.011	-
I063	*Sigara*	0.012	0.041	0.046	0.014	0.035	0.040	0.011	0.032	0.028	0.007	0.016	0.021
I064	*Gerris*	-	-	-	-	-	-	-	0.006	0.007	-	-	0.008
I081	*Calopteryx*	-	-	-	-	-	-	-	-	-	-	-	0.007
I087	*Ischnura*	0.006	0.009	0.025	0.013	0.008	0.021	0.008	0.008	0.016	-	0.005	0.010
I106	*Nemoura*	0.010	-	-	0.005	-	-	0.013	-	-	0.015	-	-
I119	Hydropsychidae	0.006	-	-	-	0.008	0.007	0.012	0.009	0.013	0.018	0.008	0.016
I120	Hydroptilidae	-	-	-	-	-	-	0.005	0.007	-	0.007	-	-
I122	Leptoceridae	0.010	-	-	0.006	-	-	0.007	0.006	0.006	-	-	-
I123	Limnephilidae	0.005	-	-	0.006	-	-	0.017	-	-	-	-	-
M06	*Dreissena*	0.011	0.005	-	0.011	-	-	0.008	0.007	0.005	0.013	0.024	0.009
M08	*Pisidium*	0.049	0.027	0.031	0.038	0.043	0.028	0.050	0.038	0.034	0.040	0.031	0.038
M09	*Sphaerium*	0.010	0.014	0.012	0.010	0.010	0.007	0.012	0.014	0.009	0.009	0.008	0.010
M12	*Bithynia*	0.015	0.018	0.013	0.023	0.017	0.019	0.019	0.022	0.020	0.015	0.023	0.019
M13	*Pseudamnicola*	-	-	-	-	-	-	-	-	-	-	0.011	0.012
M17	*Lymnaea*	0.045	0.027	0.032	0.021	0.032	0.026	0.031	0.031	0.020	0.015	0.014	0.020
M20	*Physa*	0.015	0.032	0.044	0.014	0.034	0.048	0.010	0.025	0.032	-	-	-
M21	*Physella*	-	-	-	-	-	-	-	0.008	0.012	0.013	0.023	0.046
M23	*Anisus*	0.008	0.005	0.006	0.008	0.008	0.005	0.009	0.007	0.006	-	-	-
M27	*Gyraulus*	-	0.008	0.008	0.006	0.008	0.009	0.006	0.008	0.010	-	0.008	0.012
M31	*Planorbis*	-	-	0.007	-	0.007	-	-	-	-	-	-	-
M33	*Valvata*	0.031	0.019	0.019	0.015	0.020	0.021	0.026	0.021	0.026	0.020	0.022	0.020

3.2.3.2 Relationship between diversity and environmental factors

Multiple regression models and the AIC criterion identified five variables which most of the time showed a significant effect on α diversity. The five variables included ammonium (NH_4^+), nitrite (NO_2^-), nitrate (NO_3^-) and electrical conductivity (EC) (negatively associated) and dissolved oxygen (DO) (positively associated). The global LCBD indices were also mostly affected by five variables including sinuosity, NH_4^+ and NO_2^- (negatively associated), and EC and pH (positively associated). These five variables explained a small amount of variation of global LCBD indices (Table 3.7).

Table 3.7 Coefficient correlation and adjusted R^2 of multiple regression models between α diversity, global LCBD indices and environmental variables.

Period	**Season**	**Most frequently significant variables**					
Alpha diversity		***NH_4^+***	*NO_2^-*	***EC***	*NO_3^-*	***DO***	***Adj.R^2***
D1	Spring	-0.33*	—	-0.50***	-0.30**	—	0.27
	Summer	-0.46***	—	-0.16**	—	—	0.25
	Autumn	-0.15*	—	—	—	0.22*	0.18
D2	Spring	-0.15*	-0.77**	-0.29***	—	0.23**	0.29
	Summer	-0.24***	-0.51***	-0.20***	-0.06*	0.22***	0.32
	Autumn	-0.09*	—	-0.20***	—	0.18**	0.19
D3	Spring	-0.19**	—	-0.17**	—	0.18*	0.17
	Summer	-0.20***	-0.59***	-0.24***	-0.10**	0.18**	0.24
	Autumn	-0.21***	-0.65***	-0.22***	-0.08*	—	0.30
D4	Spring	-0.39***	—	-0.14*	—	0.35**	0.31
	Summer	-0.21***	—	-0.20***	-0.10*	—	0.14
	Autumn	-0.21***	—	-0.19***	-0.09*	—	0.22
Global LCBD indices		***NH_4^+***	*NO_2^-*	***EC***	***pH***	***Sinuosity***	***Adj.R^2***
D1	Spring	—	—	—	0.0104***	—	0.10
	Summer	—	—	0.0002*	0.0016*	-0.001*	0.07
	Autumn	—	—	—	0.002**	—	0.09
D2	Spring	-0.0002*	—	—	0.0026*	-0.0016**	0.05
	Summer	-0.0001***	-0.0003***	0.0001***	0.0013***	-0.0002*	0.12
	Autumn	-0.0001**	—	0.0001**	0.0012**	-0.0006**	0.11
D3	Spring	-0.0002*	—	—	—	—	0.02
	Summer	—	-0.0003***	—	0.0007**	-0.0003***	0.06
	Autumn	—	—	0.0001***		-0.0003*	0.09
D4	Spring	—	—	—	—	—	0.06
	Summer	-0.0001*	-0.001***	—	—	-0.0006***	0.11
	Autumn	-0.0001**	-0.0006***	—	0.0008*	—	0.07

* p<0.05, ** p<0.01, *** p<0.001

When the global LCBD indices were regressed against the LCBD indices of the component communities, the LCBD indices of each component always significantly influenced the variation of the global LCBD indices (Table 3.8).

Table 3.8 Coefficient correlation of multiple regression models between the global LCBD and LCBDs of component communities in different seasons of each period.

		LCBD indices of component communities				
Period	Season	$LCBD_A$	$LCBD_C$	$LCBD_I$	$LCBD_M$	Adj.R^2
D1	Spring	0.24***	0.16***	0.33***	0.11*	0.64
	Summer	0.21***	0.09***	0.40***	0.02	0.66
	Autumn	0.15***	0.13***	0.32***	0.02	0.60
D2	Spring	0.20***	0.11***	0.38***	0.10***	0.62
	Summer	0.18***	0.11***	0.40***	0.08***	0.72
	Autumn	0.16***	0.14***	0.39***	0.11***	0.67
D3	Spring	0.12***	0.12***	0.48***	0.09***	0.66
	Summer	0.14***	0.11***	0.37***	0.06***	0.62
	Autumn	0.14***	0.17***	0.34***	0.05**	0.65
D4	Spring	0.08***	0.20***	0.39***	0.07**	0.71
	Summer	0.16***	0.16***	0.36***	0.04*	0.72
	Autumn	0.13***	0.14***	0.39***	0.08***	0.64

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

3.3 Modelling and predicting

3.3.1 Performance variation of modelling techniques applied in the LMB

The occurrence of 199 species were predicted. The overall performance of modelling techniques used (i.e. LR, RF, SVM and ANN) was significantly different when considering the three performance measures: Kappa (Friedman chi-squared, F=12.3; p=0.006), AUC (F=11.2, p=0.01) and ER (F=350, p<0.001). The highest mean Kappa (0.19) and mean AUC (0.60) were obtained for ANN, followed by LR (mean Kappa: 0.16, mean AUC: 0.59) and RF (mean Kappa: 0.12, mean AUC: 0.55), while SVM yielded the lowest mean Kappa (0.06) and mean AUC (0.53). On the other hand, a lower mean ER (0.09) was obtained for RF and SVM, while ANN and LR had a higher mean ER (0.13 and 0.16, respectively).

Based on Kappa and AUC, the performance of the models varied for different prevalence ranges (Fig. 3.8A-B). ANN and LR performed better than RF and SVM for the prevalence range <0.1; ANN, RF and LR performed better than SVM for the prevalence range 0.1-0.2; ANN and RF performed better than LR and SVM for the prevalence range 0.2-0.3, but not significantly different according to the standard error. Based on ER, RF and SVM performed better than ANN and LR in predicting species with a prevalence range <0.1, between 0.1 and 0.2 and between 0.2 and 0.3. Based on all calculations, the model performance was not significantly different for the prevalence range ≥0.3 (Fig. 3.8A).

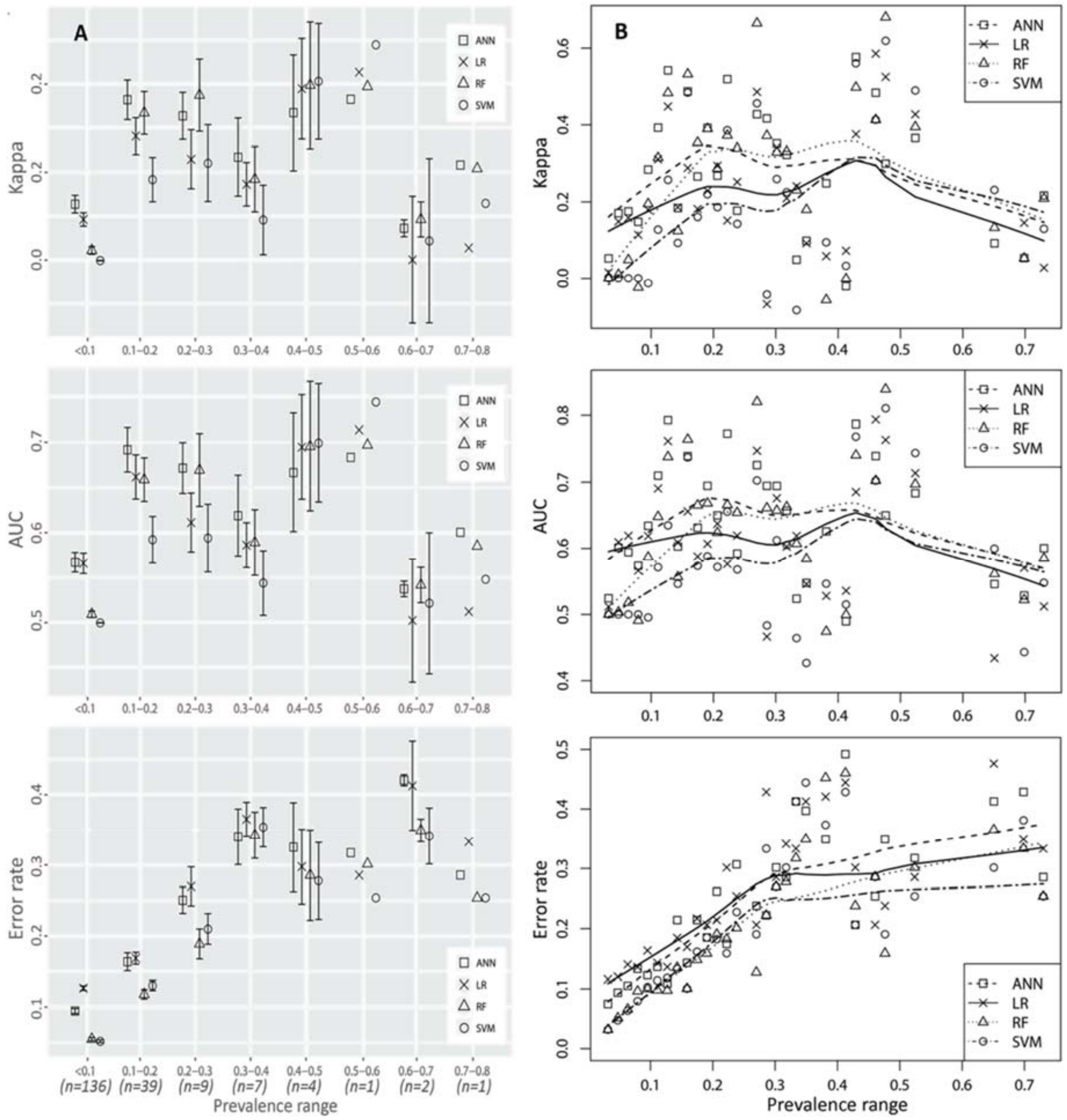


Fig. 3.8 Performance of predictive modelling techniques. A) Performance based on different prevalence ranges, B) Behaviour of the performance based on the complete prevalence range.

Linear regression models showed that the prevalence always (for each modelling technique) had a positive and significant effect on the three performance measures used. When the same data were analysed using a quadratic regression model, the explained proportion of variance of Kappa and AUC, and of ER increased, and the coefficient of the quadratic term was always negative and highly significant (p<0.001, Table 3.9).

Table 3.9 Results of linear regression models ($y=a+b_1x$) and quadratic regression models ($y=a+b_1x+b_2x^2$) for the effects of the prevalence of macroinvertebrate species on the three performance measures (Kappa, AUC and ER) of the modelling techniques. Kappa: Cohen's Kappa Statistic, AUC: area under the curve, ER: error rate, LR: logistic regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine. Asterisks indicate significance levels of regression coefficients.

Measure	Model	b_1	b_2	Adj.R^2	Measure	Model	b_1	b_2	Adj.R^2
LR					***RF***				
Kappa	Linear	0.006*		0.04	Kappa	Linear	0.012***		0.15
	Quadratic	0.613**	-0.91***	0.12		Quadratic	1.226***	-1.29***	0.31
AUC	Linear	0.003*		0.02	AUC	Linear	0.005***		0.15
	Quadratic	0.234	-0.46***	0.06		Quadratic	0.582***	-0.61***	0.31
ER	Linear	0.008***		0.61	ER	Linear	0.009***		0.73
	Quadratic	0.861***	-0.23***	0.66		Quadratic	1.008***	-0.34***	0.81
ANN					***SVM***				
Kappa	Linear	0.007**		0.04	Kappa	Linear	0.009***		0.12
	Quadratic	0.736**	-1.07***	0.12		Quadratic	0.927***	-0.72***	0.19
AUC	Linear	0.003**		0.03	AUC	Linear	0.004***		0.11
	Quadratic	0.357**	-0.56***	0.11		Quadratic	0.347***	-0.35***	0.18
ER	Linear	0.009***		0.59	ER	Linear	0.009***		0.73
	Quadratic	0.993***	-0.31***	0.66		Quadratic	1.032***	-0.40***	0.84

*p<0.05, **p<0.01, ***p<0.001.

3.3.2 Modelling alien mollusc occurrence and their co-occurrence with native molluscs

Based on CT models built for each period, the "alien" sites were not able to be correctly predicted; only the "co-occurrence" and the "native" sites could be correctly predicted (Fig. 3.9). At sites having a low sinuosity (<1.01), the co-occurrence was dependent on chemical water quality variables (e.g. NH_4^+, NO_3^-, COD, pH). Sinuosity was always one of the most important variables for the models built for each period. Based on the trees, the models revealed for each period that when the sinuosity was lower than 1.01, the co-occurrence occurred where NH_4^+ <0.4 mg/L and NO_3^- ≥2.4 mg/L for D1, COD <18.9 mg/L and NO_3^- in between 2.6-4.7 mg/L for D2, COD <21.8 mg/L and NO_3^- in between 2.3-5.1 mg/L for D3, and where pH ≥7.2 for D4 (Fig. 3.9).

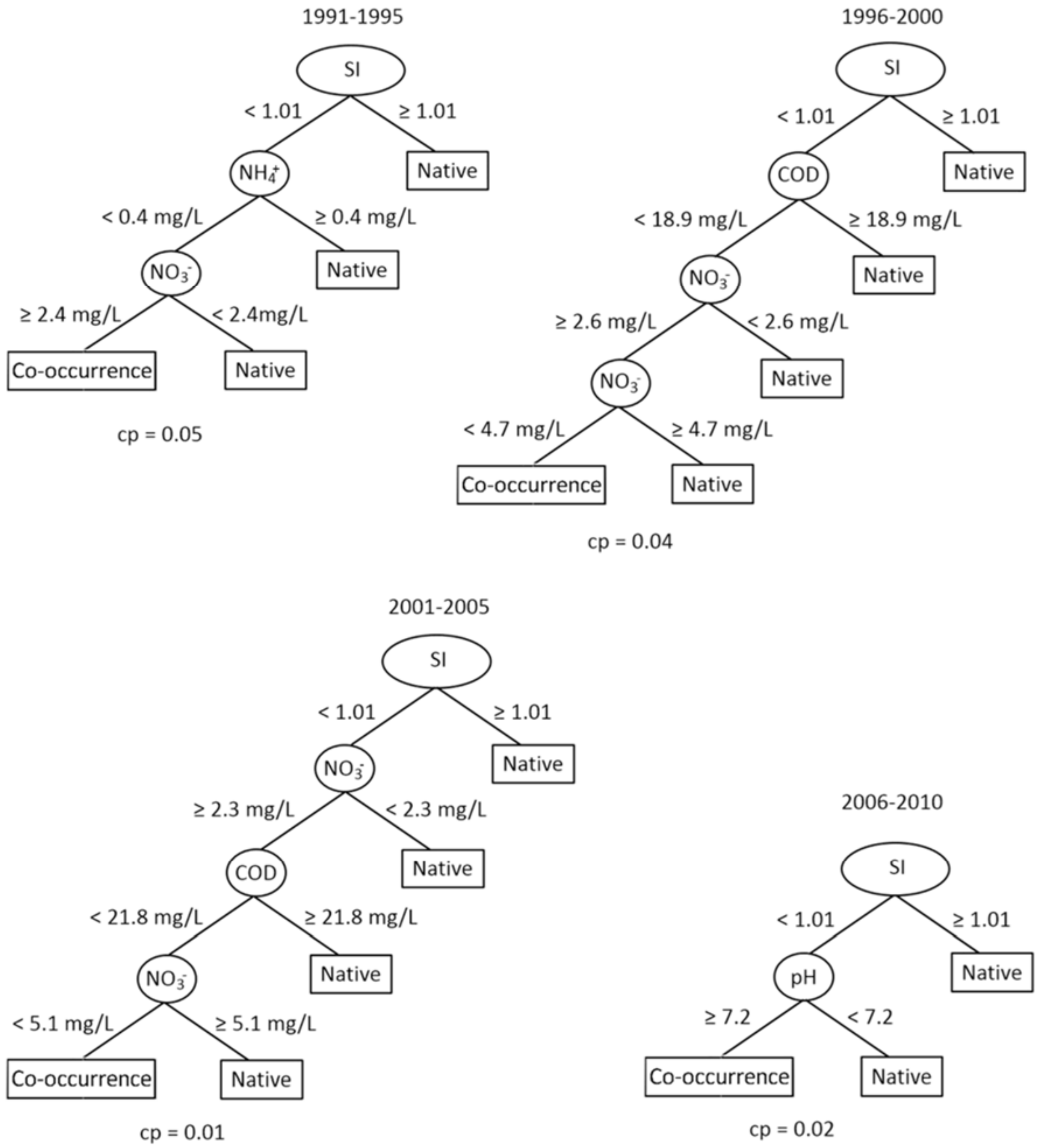


Fig. 3.9 Classification trees predicting the "co-occurrence" and the "native" sites for each period. SI: Sinuosity, NH_4^+: Ammonium, NO_3^-: Nitrate, COD: Chemical Oxygen Demand, cp: pruning complex parameter.

The mean Kappa and the mean overall CCI (Overall-CCI) of the models decreased from D1 to D4 (Fig. 3.10A-B). For the four periods (D1, D2, D3, and D4), the corresponding mean Kappa was 0.34, 0.32, 0.22 and 0.16, and the corresponding mean Overall-CCI was 93%, 92%, 79% and 54%. The mean CCI of models predicting the "co-occurrence" sites (CCI-co-occurrence) was lower for the first 3 periods (28%, 25% and 23%), while it was higher for the most recent period (52%, Fig. 3.10C). On the contrary, the mean CCI of models predicting the "native" sites

(CCI-Native) for the past 3 periods (99%, 98%, and 96%) was significantly higher compared to that of the most recent period (66%, Fig. 3.10D).

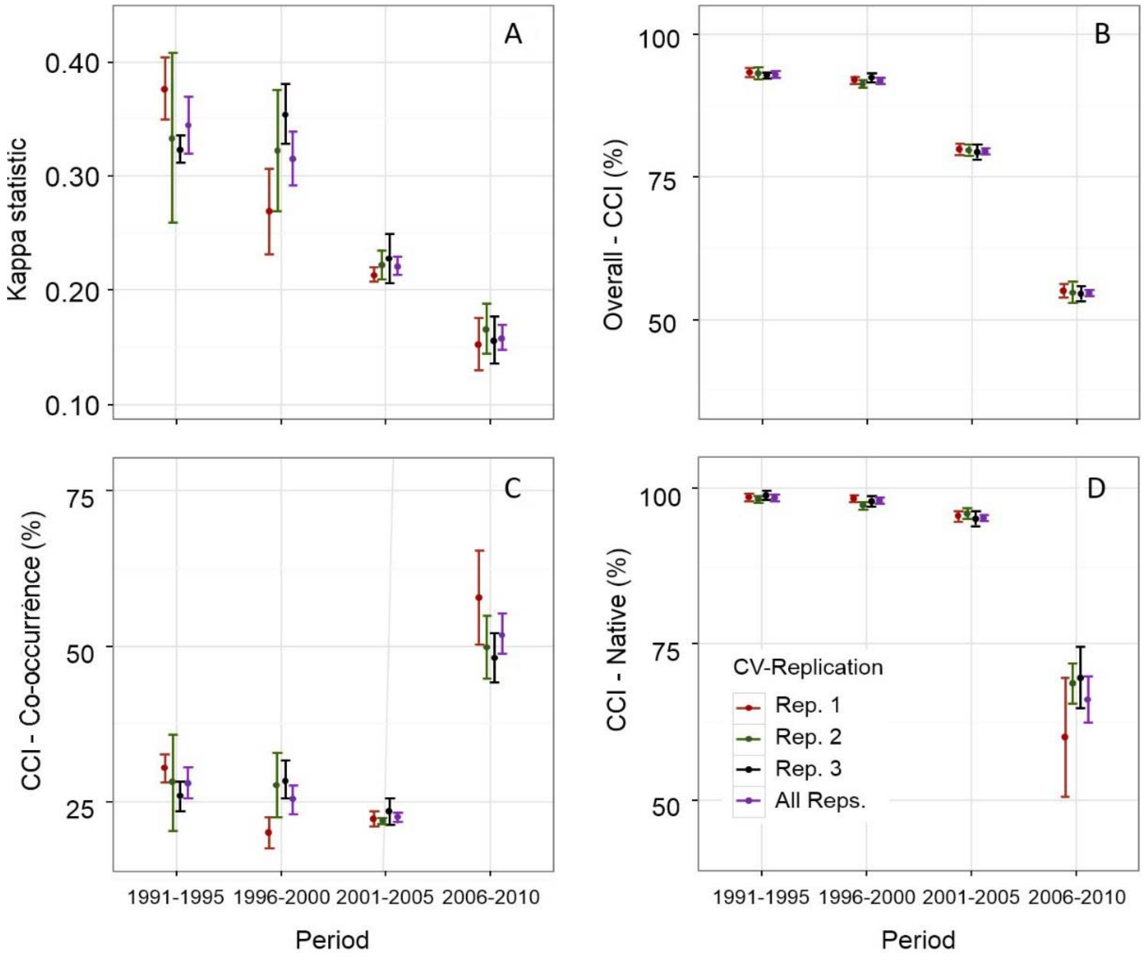


Fig. 3.10 Mean and standard error bars indicating the overall model performance based on Kappa (A) and CCI (B), and the predictive power of the predicted class "co-occurrence" (C) and the class "native" (D). CV: cross-validation, Reps: Replications.

3.3.3 Optimizing the prediction of alien mollusc occurrence

The CT models optimized by incorporating the field and cloned data (i.e. a dataset obtained by independently duplicating the field data points by *k* different individuals) were able to hindcast and forecast the class "alien" and the other two classes. The overall performance of the optimized models decreased from D4 to D1 for the hindcasting scenario, while it decreased from period D1 to D4 for the forecasting scenario (Fig. 3.11A-B). The calibrated and validated models using the two resampling approaches also yielded a similar trend in predictive performance and in stability for each scenario (see Article 5). The CCI of the class "alien" followed the decreasing trend of the overall performance for each scenario (Fig. 3.12).

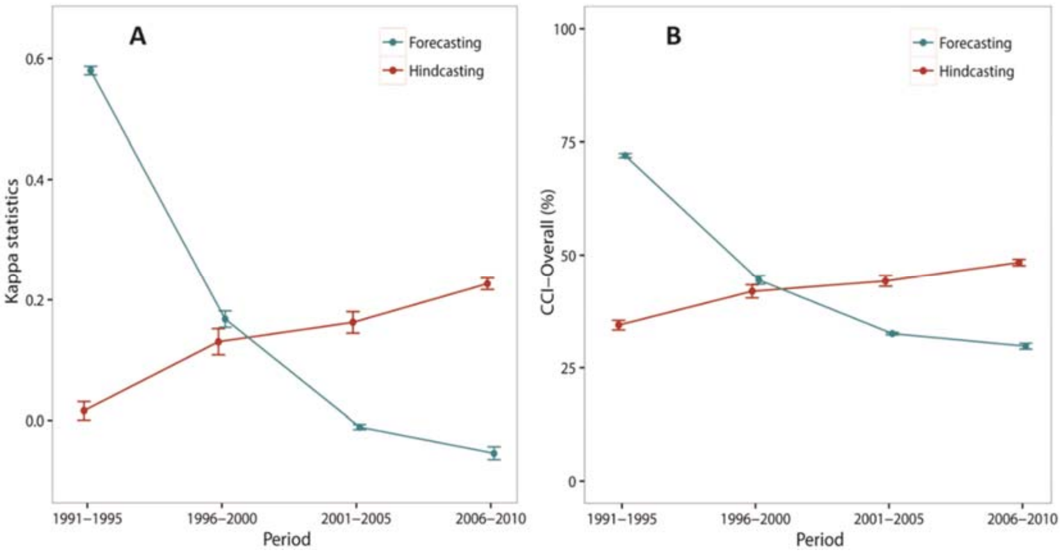


Fig. 3.11 Mean and standard error bars indicating the overall performance of CT models based on Kappa and CCI, when using the combined data and random resampling approach.

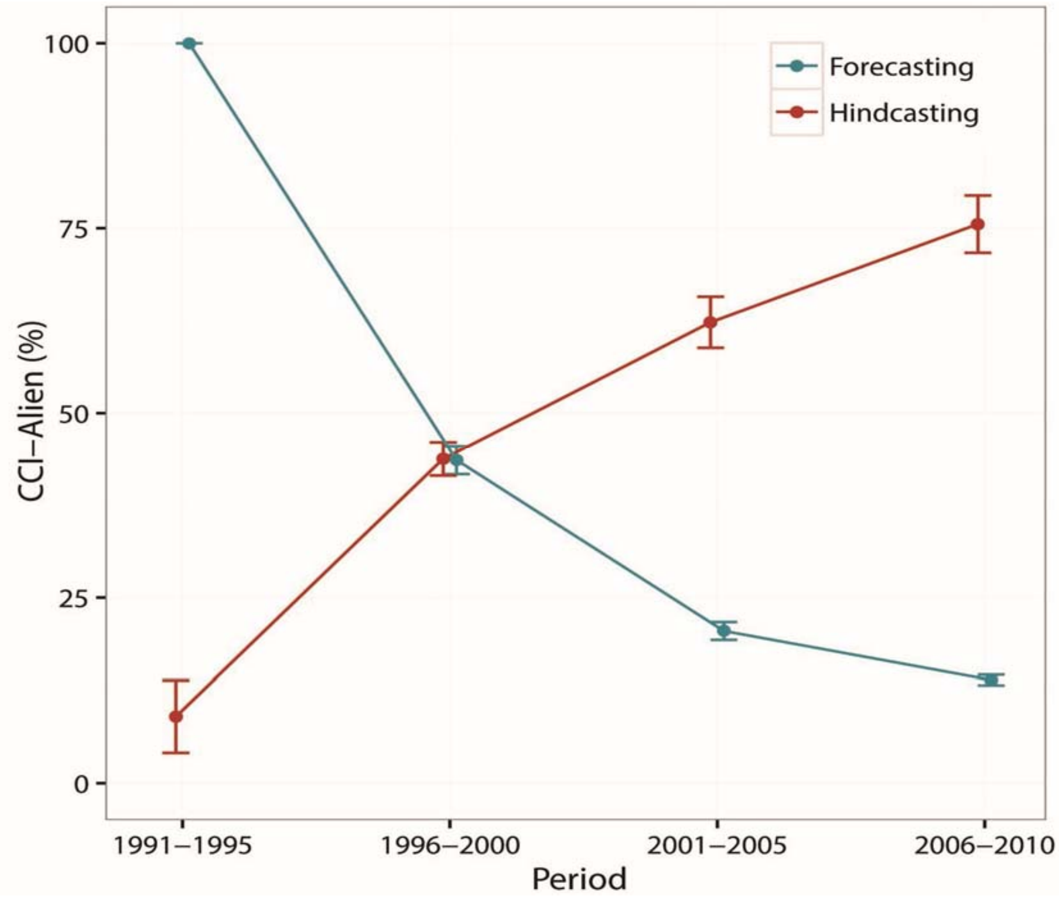


Fig. 3.12 Mean and standard error bars indicating the performance of CT models based on the correct prediction (CCI) of the predicted class “alien”.

4. General Discussion

4.1 Overall community composition and diversity in the two river systems

The “diversity” of macroinvertebrates in the LMB were found to be greater than those in Flemish rivers. This could be explained by the fact that most invertebrates from the LMB were identified to the species level, leading to a higher diversity; while the invertebrates from Flemish rivers were identified only to family or genus level. However, these findings indeed may not reveal the true composition and diversity for the LMB because it is very large, compared to Flemish rivers, but has been little studied. Thus, it is mostly likely that the number of reported taxa from the LMB is underestimated; the further it is investigated, the higher number of taxa and diversity will be. On the other hand, Flemish rivers have been extensively investigated and monitored regularly.

Regardless of the surface areas, tropical Asia has a unique geological arrangement (Sodhi et al. 2004). The evolutionary history and together with ecological interactions can explain the high biodiversity in tropical Asia (Agrawal et al. 2009). This region, as well as other tropical regions, is regarded as the ‘cradle’ and the ‘museum’ of species diversity because it is characterized by a high speciation and low extinction rates, compared the temperate regions (McKenna and Farrell 2006). Clear evidences have been shown by global diversity of many taxonomic groups including stoneflies, dragonflies, mussels, crabs and shrimps (De Grave et al. 2008; Fochetti and Tierno De Figueroa 2008; Kalkman et al. 2008; Yeo et al. 2008; Bogan 2008); a higher number of taxa (species, genera, and families) are mostly found in tropical Asia than temperate and some other tropical regions.

Based on the computation of β diversity for each component community, molluscs and insects had a higher total variation for both systems, compared to crustaceans and annelids. This could be due to the responses of different taxonomic group of each components to environmental changes and to the ecological process of each system that can support different taxonomic groups. For example, mayflies and stoneflies (i.e. genera) have been found to have a high similarity between temperate Europe and tropical Asia, compared to some other geographic regions (Barber-James et al. 2008; De Moor and Ivanov 2008). A high gastropod species diversity have been reported from Europe (e.g. Southern France and Spain, Southern Alps and Balkans regions) and tropical Asia (the LMB) (Strong et al. 2008). This may indicate high β diversity of insects and molluscs is favoured by ecological and environmental conditions from both regions which promote their distribution and abundance, and thus lead to a higher β diversity. A more detailed discussion on the differences and on responses to environmental variables is provided separately in the following paragraphs for each system.

4.2 Spatio-temporal changes of communities and their relation to environmental factors

4.2.1 The LMB

4.2.1.1 Spatial organization and environmental factors

The compositional and diversity differences found for the four macroinvertebrate assemblages were related to spatial environmental heterogeneity observed from the upstream to the downstream parts of the LMB (Dobson et al. 2002; Heino et al. 2005; Wang et al. 2012; Salmah et al. 2014). This can be explained by the proportion of the richness, abundance, diversity and the represented indicator taxa of the different taxonomic groups for each assemblage. Although assemblage IIa, IIb1 and IIb2 were characterized by a low and similar macroinvertebrate richness, abundance and diversity, each component (i.e. annelids, crustaceans, molluscs and insects) contributed differently to the overall macroinvertebrate community composition for the particular assemblages. On the other hand, assemblage I had a high richness, abundance, diversity and a high number of indicator taxa. Therefore, these findings clearly suggested that the community composition of the four assemblages, particularly assemblage I, is distinctly organized and structured along the environmental gradients of the LMB.

Site-specific characteristics, connection in habitats and mesohabitats and variation in river morphology can strongly affect community structure and composition found from the upstream to the downstream parts of rivers (Thorp and Delong 1994; Zilli and Marchese 2011; Mazão and Bispo 2016). Among the measured environmental variables, physical conditions and land cover types are the key factors determining macroinvertebrate assemblages along the LMB. The surface area of watershed, river depth and altitude are the most important variables discriminating the upstream assemblages (IIa and IIb2) from the downstream assemblage (I). This discrimination is clearly explained by the axis 1 of the LDA accounting for 65.7 % of the total variance of the data and by a high predictive reliability for each particular assemblage (Fig. 3.3A-B). Because of the observed physical conditions, together with other related variables (e.g. water temperature, dissolved oxygen and land cover types), the assemblages were characterized by a very different community structure, composition and diversity.

The intermediate assemblage (IIb1) in-between was positively linked to three types of land cover: agricultural land, inundated and wetland. Along axis 3 of the LDA ordination, these land cover types are the key factors best explaining the macroinvertebrate composition of assemblage IIb1 (Fig. 3.3C-D). However, the proportion of inundated and wetland constituted a small percentage of the land cover, and together with agricultural land, they only explained a limited amount of the variance of the data for axis 3 of the LDA model (11.2%). As a result, the predictive performance for this assemblage was not as high as the performance for the other assemblages (see Sor et al., 2017a). A broad range of values of physical-chemical variables (e.g. dissolved oxygen, river width and depth, the surface area of watershed and altitude) found at sites belonging to this assemblage could also explain the performance yielded. On the other hand, assemblage IIa was associated with two important factors (electrical conductivity and urban areas) according to the LDA axis 2, which explained a higher variance of the data (23.1%). Therefore, these two factors, together with other measured environmental variables, could correctly predict this assemblage with a higher reliability. The possible explanation for the association found is that many sites that belong to this assemblage are located along the tributaries (e.g. the Mun and Chi river basins in Thailand) and the main channel where cities were built. The tributaries are surrounded by intensified agriculture and the cities are exposed to a high level of anthropogenic disturbance (Dao et al. 2010; Kudthalang and Thanee 2010). Therefore, the runoff from the surrounding agricultural areas and the discharge of urban wastewaters may cause the increase in electrical conductivity (Wetzel 2001).

4.2.1.2 Spatial diversity pattern and environmental factors

The distinct organization of community composition and α diversity from the up to the downstream leads to a high β diversity. Alpha diversity pattern follows the richness and abundance patterns of each assemblage; a high diversity was observed at the downstream sites, which are characterized by a large watershed surface area, deep and wide rivers and a high water temperature, and a low diversity found for the upstream and tributary sites, which are characterized by highland area, clear water, high DO concentration and dense forest cover.

The most important taxa, which mainly contributed to the high β diversity, found at the downstream were dominated by annelids, crustaceans and molluscs. This is because they are more abundant in the downstream part of the LMB, as in the case of two mollusc species: *Corbicula leviuscula* and *Limnoperna siamensis*. Unsurprisingly, this result supports previous studies (Arscott et al. 2005; Collier and Lill 2008; Wang et al. 2012). On the other hand, at the upstream sites, most of the important taxa belong to insects. This could be due to that fact that clear water and high values of dissolved oxygen, which are mostly found in tributaries and upstream sites, are mainly preferred by insect taxa (Dobson et al. 2002; Collier and Lill 2008; Królak and Korycińska 2008). For instance, the two important taxa found with the highest

SCBD indices, *Polypedilum* sp. (I033) and *Ablabesmyia* sp. (I017) were highly associated with sites having high values of SD, which were mostly in tributaries (see Sor et al., 2017a).

The β diversity of macroinvertebrates was estimated as the total variance (BD_{Total}) of the communities found at the sampling sites and computed the contributions of individual sampling sites (LCBD indices) to BD_{Total}. LCBD values are scaled to add up to 1 over the whole study; the mean LCBD value in the LMB was thus 1/63 = 0.016. The variation in LCBD indices (range: 0.010–0.023) appeared to be related to the environmental factors that mainly explained the community composition and α diversity. LCBD indices were positively associated with river depth, surface area of watersheds and electrical conductivity. High values of these variables are most of the characteristics of sites situated in the delta and along the main channel of the LMB (Sor et al. 2017a). This indicates that a high degree of uniqueness in composition of benthic macroinvertebrates in this tropical system mostly occurred along the mainstream of the rivers, which are highly associated with anthropic disturbance (see Article 2). Small LCBD indices were linked to the tributary sites, which are characterized by a high values of Secchi depth (rivers with clear water). Due to this fact, tributaries mainly support particular assemblages of particular macroinvertebrate taxa, e.g. Ephemeroptera, Plecoptera, Trichoptera and Diptera (Dobson et al. 2002; Wang et al. 2012).

When regressed with LCBD indices of component communities, the global LCBD indices was highly linked to the taxonomic compositions of the mollusc and insect communities ($LCBD_M$ + $LCBD_I$ indices). However, the degree of uniqueness in composition of macroinvertebrate communities (global LCBD indices) is expected to be contributed by the component communities. In the LMB, LCBDs of mollusc and insect communities, which had a higher total variation ($BD_{MTotal} = 0.78$ and $BD_{ITotal} = 0.74$, respectively), explained most of the global LCBD variation because these two groups had higher abundances and wider distributions than the annelid and crustacean communities, which had lower total variation ($BD_{ATotal} = 0.72$ and $BD_{CTotal} = 0.38$, respectively). De'ath (2002) and Davidson et al. (2010) explained that taxa with low richness and low occurrence lead to a less variance of the community composition, and this is similar to the present findings for the annelid and crustacean communities.

4.2.2 Flemish rivers

4.2.2.1 Spatio-temporal changes in composition and environmental factors

The benthic community composition in Flemish river system was found to gradually change (e.g. macroinvertebrate richness and abundance) from the early 1990s to the late 2000s. The changes can be seen from the RDA plots (Fig. 4.1A-B). The circular arrangement of the monitoring periods is due to the gradual changes in species composition, and thus some taxa highly contributed to the variation in composition in the past periods, but not in the recent periods, and vice-versa (Table 3.6, Fig. 4.1B). However, the seasonal changes in recent periods (2000s) were not significantly different, except for the past periods (1990s). This could also mean that the water quality condition across seasons in each period of 2000s is more or less the same, and that allows similar composition of macroinvertebrates to occupy the habitats.

The temporal changes over the four periods observed corresponded to the improvement of water quality condition from the past to the recent period (Leuven et al. 2009). Clearly, the centroids of community composition in the 1990s were linked to a high value of ammonium, nitrite, total phosphorus, and COD, which indicate a poor water quality. Whereas the centroid of community composition in the early 2000s was more related to a high value of DO, and thus the communities was composed of many moderate-to-high pollution sensitive insect taxa, e.g. Simuliidae (I030), Hydropsychidae (I119), Baetis (I037) (Gabriels et al. 2010). In the late 2000s, the community centroid was more related to a high value of pH and EC. This could be due to the fact that many samples from this period were mainly recorded from brackish polder watercourses and from the main harbour watercourses, e.g. river basin Ijzer, Brugse Polders (Bru-Pol), Gentse Kanalen (Gen-Kan) and Beneden-Schelde (Ben-Sch) (Fig. 4.1B) where a high level of seawater intrusion occurs and a high intensity of human-related activities takes place.

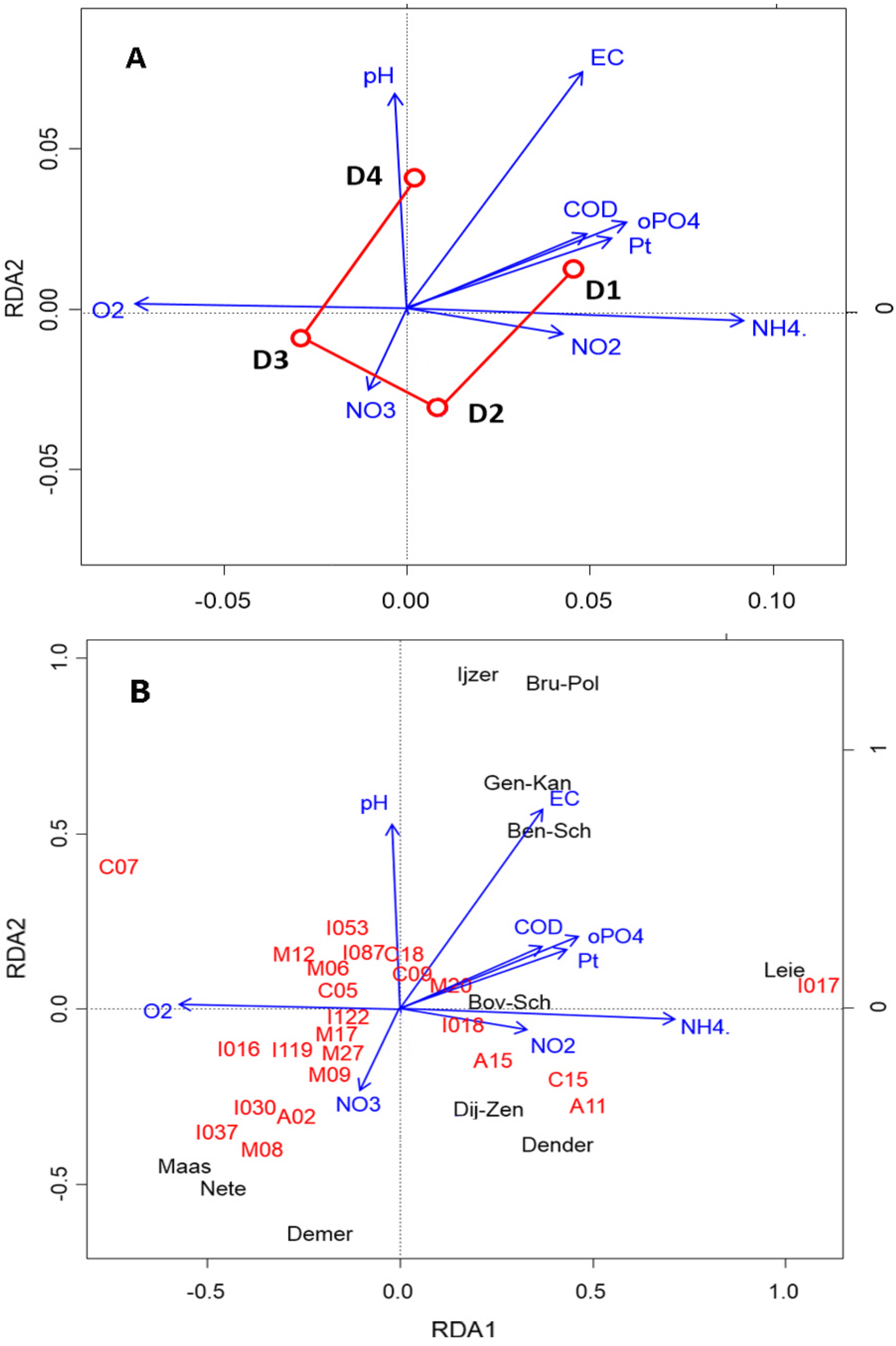


Fig. 4.1 RDA plots showing the association of (A) the centroid of macroinvertebrate communities (points) with environmental variables in each period and of (B) the relationship between different taxa, environmental variables and river basins: Ijzer, Leie, Maas, Nete, Demer, Dender, Bru-Pol: Brugse Polders, Gen-Kan: Gentse Kanalen, Ben-Sch: Beneden-Schelde, Bov-Sch: Boven-Schelde, and Dij-Zen: Dijle Zenne. Taxa (red colour) that had small loading scores on the RDA axes were removed to improve visibility; see Table 3.5 for full name of the taxa.

4.2.2.2 Spatio-temporal diversity patterns and environmental factors

Alpha diversity of macroinvertebrates was undoubtedly related to the improvement of water quality in Flanders. A gradual increase in α diversity from the early 1990s to the early 2000s could be explained by the gradual increase in the overall number of taxa and abundance of all macroinvertebrates and of each component (Fig. 3.4, Table 3.4). However, the number of taxa of the global communities in the late 2000s slightly decreased or remained more or less stable, compared to the early 2000s, but the their abundance spectacularly increased. This pattern was also observed for the four component communities, of which insects and molluscs had a higher α diversity, compared to annelids and crustaceans (Fig. 3.6). This could be due to the dominance of some taxa (e.g. alien taxa), which highly benefit from the water quality improvement, and thus increases the homogenization of the communities and subsequently reduces the diversity (Rahel 2002; Rahel 2007). A clear evidence can be observed from the dominance of an alien mollusc *Physella* over the native molluscs *Physa* in Flemish rivers in the recent periods (see Appendix T4). The seasonal differences in α diversity were found only for the 1990s (significantly low in Spring), but not the recent periods (2000s), as found for richness and abundance explained earlier.

Similar to other studies, α diversity of the animals in Flemish rivers is affected by nitrate content (e.g. ammonium, nitrite, nitrate), electrical conductivity and dissolved oxygen (Mereta et al. 2012; Md Rawi et al. 2013; Suhaila and Che Salmah 2014). The higher concentration of nitrate content and electrical conductivity, the more it influences the living conditions of the benthic communities (Friberg et al. 2010; Boets et al. 2013), and thus negatively affects α diversity, as found in the present study (Table 3.7). A positive association between α diversity and dissolved oxygen was also observed, and this is generally the case found in previous investigations (Baptista et al. 2001; Md Rawi et al. 2013; Suhaila and Che Salmah 2014).

Beta diversity of the global communities across sampling sites in Flemish rivers was relatively high (range: 0.49-0.60). A higher β diversity was found for the recent periods, compared to the past. The same pattern was found for β diversity of each component (Table 3.5). This indicates that the community composition varies a lot from the past periods (1990s), when the water quality was very poor, to the recent periods (2000s), when the water quality was substantially improved. The water quality improvement also influenced the seasonal β diversity of global communities in each period. Similar to α diversity patterns, a lower β diversity was found in Spring, compared to Summer and Autumn, during the 1990s, but not during the 2000s.

The five most important taxa (i.e. Naididae (A11), Gammaridae (C07), Asellidae (C15), Chironomidae non *thummi-plumosus* (I016) and Chironomidae *thummi-plumosus* (I017)) were very widespread and abundant (see Appendix T4) over the four periods. This is the reason that these five taxa greatly contributed (between 5%-14%) to β diversity in each period. However, Naididae (A11), Asellidae (C15) and Chironomidae *thummi-plumosus* (I017) were closely linked to the centroid of site distribution in the early and late 1990s (Fig. 4.1B), suggesting that they are able to withstand the poor water quality (Feld and Hering 2007; Gabriels et al. 2010). The other taxa that were responsible for the temporal changes in community composition include Enchytraeidae (A08), Haplotaxidae (A09) and other taxa (e.g. *Bythinella,* Blephariceridae and *Arctocorisa*), which occurred in the 1990s, but disappeared in the 2000s (Table 3.6, Appendix T4). Moreover, the different composition of communities is also caused by a wide distribution and high abundance of Dryopidae, *Gerris*, *Calopteryx*, *Physella,* and the present of new taxa (e.g. *Menetus* (an alien mollusc)*, Panopeidae* (crustacean), *Somatochlora* and *Procloeon* (insect)) in the 2000s (Table 3.6, Appendix T4).

The β diversity among sites for each period was estimated from the LCBD indices of corresponding period. In most cases, the LCBD values were positively associated with pH and EC (as found in the LMB), while negatively associated with nitrate content (ammonium and nitrite) and sinuosity. The association found suggests that sites with a high degree of uniqueness (i.e. high LCBD values), which may indicate high or low species richness as shown in Legendre and De Cáceres (2013), mainly occur in straight rivers that have low concentration of nitrogen content, but high concentration of EC and pH. This seems to be in the case of sites situated in the main harbour watercourses and in brackish polder where rivers have a low sinuosity, but have high values of conductivity and pH (Fig. 4.1B). The physical-chemical conditions of these watercourses may reduce taxonomic composition. Only some taxa belonging to Oligochaeta, Chironomidae (Diptera) and Physidae (Molluscs), which are pollution tolerant and can occur in areas with high pH and EC values (Rodrigues Capítulo et al. 2001; Feld and Hering 2007; De Troyer et al. 2016), remain, and thus taxonomic composition may become unique (high LCBD indices), but not more diverse, by having their number of taxa reduced.

The LCBD indices of global communities were found to be highly contributed by the four component communities' LCBDs, as reflected by their significant correlation (Table 3.8). This result is somewhat different from that found for the LMB; only molluscs and insects that were highly and significantly contributed to the global LCBD values. However, the finding for

Flemish rivers is more logic because, as mentioned earlier, the global LCBD indices is expected to be contributed by the LCBDs of each component community.

4.3 Model development, performance and predictions

4.3.1 Modelling techniques and their application in the LMB

According to Kappa and AUC, which have been shown to be a better measure compared to error rate that is highly biased (Fielding and Bell 1997; Manel et al. 1999; Sor et al. 2017b), ANN performed the best across the complete prevalence range and a lower prevalence range (i.e. <0.1, in the case of rare species), compared to other techniques (Fig. 3.8). Some authors (e.g. Mastrorillo et al., 1997; Pearson et al., 2002; Segurado et al., 2004) agree that ANN provides advantages over other techniques for predicting species occurrence, whilst others found RF to be better (Grenouillet et al. 2011; Gallardo and Aldridge 2013). This could be due to different input predictors, parameterization settings and the applicability of each technique. Setting the complex parameter of the models has to be carefully taken into account because they may lead to an over-fitted model, which produces a less general result with a lower applicability to different situations (Babyak 2004). Model applicability is dependent on the type of data (e.g. missing values and data distribution, De'ath and Fabricius, 2000; Therneau and Atkinson, 1997) and on where the data is derived (e.g. different ecological regions, Guisan and Thuiller, 2005; Randin et al., 2006). Most of the studies evaluate and apply modelling techniques based on the data collected within particular regions, but the applicability of those techniques to a different geographic range is hardly assessed (Fielding and Haworth 1995; Özesmi and Mitsch 1997; Kleyer 2002; Everaert et al. 2014; Forio et al. 2016). As such, the varied performance of modelling techniques may reveal a unique behaviour of each technique and its suitable applicability (e.g. ANN) for the Lower Mekong Basin and the neighbouring areas rather than for other geographic areas.

The performance of modelling techniques was highly dependent on species prevalence (Table 3.9), suggesting that each technique may have different behaviour when predicting species with different prevalence range. For all models, the best performance was found for common species with an intermediate prevalence (e.g. 0.4-0.6, Fig. 3.8). In between this prevalence range, there is a smaller bias for the models to select presence/absence data for training and validation sets. A more or less equal distribution between the presence and absence, which are likely to result in many correctly predicted instances of both the true positive and true negative fractions, could be the main reason responsible for the best predictive performance (Manel et al. 2001; Allouche et al. 2006). For very common/generalist (prevalence ≥0.6) and for rare species (prevalence <0.1), the models yielded a low performance. In these cases, there is a high imbalance between the presence and absence data. Therefore, the models can correctly predict many true positive instances and a few or perhaps no true negative instances for common species, and vice versa for rare species. This explains the low performance of the models because the measures (i.e. Kappa and AUC) are designed to reflect model performance in absence and presence instances simultaneously (Cohen 1960; Cicchetti and Feinstein 1990; Zweig and Campbell 1993). Thus, a few or no instances of either the true positive or the true negative fraction results in a low performance.

4.3.2 Predicting alien species occurrence and their co-occurrence with native molluscs

Based on a set of chemical water quality variables, the CT models were able to correctly predict the co-occurrence alien and native molluscs, and the occurrence of native molluscs; but not the occurrence of the alien alone. Sinuosity was always one of the most important determining factors because it formed the main root in all models. Sites having a low sinuosity (<1.01), which mainly corresponds to straight rivers, may be subjected to a high number of passing ships, which is considered the main pathway of invasions (Bij de Vaate et al. 2002; Nunes et al. 2015). Moreover, straight rivers shorten travelling distances, resulting in more frequent transportations and thus allow a large and frequent amount of ballast water being released. Consequently, due to a higher number of introductions, the survival rate of alien molluscs increases (Gollasch 2006). The hotspots of mollusc invasion (the alien and the co-occurrence sites) were mainly situated in brackish polder watercourses and at large rivers that have a shorter distance to the ports in the Rhine delta, to the coast of the North Sea and to the other large rivers (e.g. Meuse River). This observation is also supported by evidence in Boets et al. (2016) and in Grabowski et al. (2009) who found that alien fauna mostly inhabited large rivers where intensive navigation takes place. The followed key variables responsible for the co-occurrence included NH_4^+, NO_3^-, COD and pH. These variables are the major factors influencing the distribution and diversity of freshwater fauna (Leuven et al. 2009; Wang et al. 2012). Thus, for each period, where sampling sites had a low sinuosity (<1.01), the co-occurrence was mainly dependent on the chemical water quality status (see Article 4).

The overall performance of the models (Kappa and Overall-CCI) in predicting the co-occurrence was moderate to good. When using data from the past period, the performance was higher than when using data from the most recent period (Fig. 3.10A-B). This fact could be attributed to less complex biotic interactions and to the limited number of occurrences of alien species in the past. For example, most alien and native species had a low occurrence and abundance in the past periods (Appendix T4). In this context, alien species may invade those sites where competition is low and where a few native species occur. Some alien taxa have a strong ability to compete for food and niche with the native species. For examples, in the past, the alien *Physella* occurred at only few sites with a few individuals, but then widely spread with high abundance in the recent periods, leading to a reduced abundance of the native *Physa* (Appendix T4). This could be a result of competition for food and niche with the alien *Physella*. Araújo and Luoto (2007) and Meier et al. (2010) stated that the biotic interaction is important in predicting species distribution, and in general when included in the model, it increases the predictive performance. Provided that information on biotic interaction was not included in the models and that the alien molluscs already occupy a wide range of environmental conditions in the most recent data, the overall performance of the models based on the most recent period is somewhat lower. Moreover, during the continuing expansion phase of invasion (the late 2000s), the range of environmental conditions where alien species occur increased. At the same time, many native species recovered their occurrences and abundance, due to improved water quality. These findings might also explain the higher co-occurrence and thus the lower overall performance of the models.

However, the prediction of the "co-occurrence" of alien and native molluscs was more reliable for the recent period compared to the first three periods. This could be explained by the increased co-occurrence sites observed in the most recent period (see Article 4). This is similar to the models predicting the occurrence of native molluscs. The predictive models yielded a high performance when the prediction was based on the periods (i.e. the first three periods) that have a large sample size of the "native" sites. This is quite logic as for predictive models, the more input samples provided, the better the models learn and as a result, a higher predictive performance can be obtained (Stockwell and Peterson 2002; Hernandez et al. 2006).

4.3.3 Optimizing the prediction of alien mollusc occurrence

Since the CT models were not able to predict the sole occurrence of alien mollusc, which could be due to a low number of instances of this particular class, the CT models were optimized via a hindcasting and forecasting approach. Based on the best CT model configuration (for both in terms of performance and interpretation) obtained, a similar sample size of each class appears to be the best input, and thus should be considered in every predictive model as has been generally suggested (e.g. Manel et al. 2001; Allouche et al. 2006). A clear evidence was revealed from the CT models that used the field data (field observations). For these models, the prevalence of each predicted class was not equally distributed and the models were unable to predict the putative occurrence of the alien molluscs. However, by using the combined data, their occurrence could be correctly predicted at past periods (D1, D2, and D3) when the CT models were calibrated based on the data of the most recent period (D4) and at the recent periods (D2, D3 and D4) when the models were calibrated based on the data of D1. These results correspond to the field past and current situation.

One of the challenges in predicting a species' occurrence is obtaining a balanced class distribution for the response variable, because a small sample size of each predicted class can be the source of instability and errors in species distribution models (McPherson et al. 2004; Allouche et al. 2006). However, results from this case study demonstrated that having the exact same class distribution (i.e. stratified split) is not always necessary during the calibration and validation process because the models can make the correct prediction and yield a similar performance when the prevalence reaches a certain threshold. However, the model is incapable of predicting a class with a very low prevalence, as was the case of the class "alien" in this study. As collecting new field data from past/recent periods is unfeasible/costly, optimization appears necessary (Hirzel and Guisan 2002). In respond to this, I optimized the prediction by considering cloned data for the models, resulting in a better prediction and higher reliability. Cloned data have been applied and recently suggested for hierarchical models in ecology (Lele et al. 2007; Ponciano et al. 2009; Lele et al. 2010). The increasing number of clones used in the models can increase the maximum likelihood of the prior class distribution, but it does not affect the statistical accuracy, which mainly depends on the information of the field data and the model calibration and validation process (Lele et al. 2007). This suggests that CT models can be successfully optimized and improved to predict the actual occurrence of alien molluscs by incorporating cloned data into the models.

5. General Conclusion and Perspectives

5.1 General conclusion

The different macroinvertebrate community composition, which leads to a different amount of variation (i.e. β diversity), is mostly likely driven by environmental conditions characterizing each system. A higher diversity is found for the LMB, compared to Flemish rivers. Identification effort could be a reason causing the differences; most benthic animals from the LMB were identified to the species level, whilst those from Flemish rivers were identified only to family or genus level. However, these results tend to be revealing only a portion of knowledge on macroinvertebrates for the LMB, because this basin is very large, but has been seldom studied. Thus, the number of reported taxa is most likely to be underestimated; the further investigations will greatly increase the systematic and ecological knowledge on macroinvertebrate from this basin. On the other hand, Flemish rivers have been extensively investigated and monitored regularly over the past decades. Nevertheless, the two systems showed some similarities; molluscs and insects had a higher total variation and highly contributed to the total variation of the global communities, compared to crustaceans and annelids. Moreover, a high degree of uniqueness in community composition (i.e. high LCBD values) frequently occurred at sites where a higher level of anthropic disturbance was observed.

In the LMB, the most important variables influencing the macroinvertebrate composition and diversity include altitude, surface area of watersheds, river width and depth, Secchi depth, DO, EC and water temperature. Together with the highly variable topography, geology, hydrology and different land cover, the composition and diversity of macroinvertebrates in the LMB is spatially and distinctly organized from the up to the downstream of the basin.

In Flanders, river morphology (e.g. sinuosity) and physical-chemical water quality variables (e.g. DO, pH, NH_4^+, NO_3^-, NO_2^-, COD and EC) are the key variables. Water quality improvement from the past to the recent period had gradually enhanced the composition and diversity of benthic animals, leading to an increased β diversity in the recent period. The improvement of water quality also promoted macroinvertebrate composition and diversity, which significantly varied among seasons (i.e. Spring vs Summer vs Autumn) in the past 1990s, to be seasonally similar in the recent periods (2000s). However, good water quality appeared to drive community homogenization. This is clearly indicated by a decreased local (α) diversity in the 2000s.

Among modelling techniques applied in the LMB, ANN performed the best across the complete species prevalence range. ANN also yielded the highest performance when predicting the occurrence of rare species. In Flemish rivers, the CT models could only predict the co-occurrence of alien and native molluscs, and the occurrence of native molluscs; but not the sole occurrence of alien species. Via a hindcast- and forecasting approach, the CT models were successfully optimized to reliably forecast and hindcast the sole occurrence of alien molluscs, by incorporating cloned observations into the models. Whether in the past or recent periods, the results of this optimization correspond to field observations.

5.2 Implications for management and restoration

Diversity of global and component communities is biologically and ecologically important for ecosystem structure and functioning. Thus, a sudden change in the composition of any taxonomic group may result in disproportionate or unexpected responses of other taxa in the system (Naeem 1998), and consequently alter ecological processes (Covich et al. 1999). In this context, the indicator species, the species/taxa with high among-sites variance (high SCBD indices) and the uniqueness in community composition at sampling sites (LCBD indices) identified/analysed in this study have provided insights into the ecological importance and environmental degradations of sampled sites/river reaches (see Sor et al. 2017a, Article 2). With these results, together with the key correlated environmental factors and the implemented modelling techniques, habitat quality and suitability for the most vulnerable native and pollution sensitive taxa could be more conveniently monitored and analysed. This knowledge could provide an advantage to support decision-making concerning conservation, management and restoration planning of these keystone taxa and their communities at local and regional scales.

With the challenge of increasing rate of invasion, which has become a major concern for global economy and environment (Sala et al. 2000), numerous threats have put high pressure on native biodiversity. Although some alien species (e.g. *Dreissena* mussels and *Corbicula* Asian clams, the well-known exotic species for European and American river systems), also play a key role in ecological functioning, e.g. improving water quality via filtering process (Higgins et al. 2011; Sousa et al. 2014), their wide spread and high abundance substantially affect other species from the local to ecosystem levels. In the case of *Dreissena* and *Corbicula,* they compete with native species, reduce plankton communities, cause a decline of dissolved oxygen and transfer organic resource from the water column to the sediments (Caraco et al. 2000; Descy et al. 2003; Caraco et al. 2006; Vaughn and Spooner 2006). Moreover, once they have successfully colonized new habitats, eradication is hardly possible (Regan et al. 2006). This consequently leads to native species replacement, food web reorganization and community composition simplification (Gurevitch and Padilla 2004; Bernauer and Jansen 2006; Didham et al. 2007). Due to these facts, applicable risk assessment approaches are required to mitigate ecological and socioeconomic impacts posted by alien species, and to develop sound prevention and management options (Panov et al. 2009). In this study, an assessment (i.e. the case in Flemish rivers) using predictive models including a hindcasting and forecasting approach provides more insights into the ecology of alien species (e.g. the environmental conditions where alien species solely exist and co-exist with native species), which can serve as a basis for invasion control and be used to support environmental management and conservation planning. However, shall be this approach or other novel assessment tools implemented, both negative and positive effects should be taken into account, because as mention earlier, alien species can also be a good ecosystem engineers or a food sources for other species (Palmer et al. 1997; Chowdhury et al. 2016).

5.3 Perspectives

Information on and insights into benthic macroinvertebrate ecology and its implications for management and restoration planning provided by this study are of great significance contributing to increase scientific knowledge of the two basins, particularly the LMB which has been seldom studied. However, there is much more to be investigated. For both systems, a deeper investigation of each component community should be conducted as they differently respond to environmental conditions. It is also recommended for further rigorous examination of temporal changes at some specific sites exposed to different anthropic disturbance or have different land cover characteristics. Between these sites, temporal species composition, variations, taxonomic and functional diversity, biological traits and the degree of uniqueness in community composition (i.e. LCBD indices) should be analysed rigorously. These can provide useful and fundamental knowledge on how the communities will in the future change under different environmental conditions (e.g. from a less to a substantial disturbed sites/locations) (Legendre and Salvat 2015), and thus a management or restoration planning can be foreseen. Moreover, should there be new data collection, other important variables including nutrients, sediment loads and habitat variables (bed rock, mud, leave litters etc.) and climatic variables (for large spatial scale like the LMB) have to be taken into account because they greatly influence macroinvertebrate community composition in both the tropical and temperate systems (Nicola et al. 2010; Cai et al. 2012; Pearson 2014).

Regarding the predictive modelling approach, an ensemble modelling evaluated with an array of performance measures (e.g. Kappa statistics, true skill statistics, ROC curve, error rate or the correctly classified instances) is recommended when comparing the performance of different techniques or when forecast or hindcast species distribution (Araújo and New 2007; Grenouillet et al. 2011). Moreover, incorporating the missed variables including those mentioned in an earlier paragraph, biotic interactions and dispersal vectors (as in the case of alien species) into the models is also suggested. This incorporation will improve the predictive performance and reliability of the models (Araújo and Luoto 2007; Boets et al. 2014; Parravicini et al. 2015). Lastly, testing model transferability into a new geographic region (e.g. outside the region where the models were calibrated) is vital to validate the model applicability. This is suggested because most of studies have evaluated and applied modelling techniques based on the data collected within particular regions, but have not access their applicability outside the studied areas. This case also applies to the present study; the CT models were successful optimized to hindcast and forecast the occurrence of alien species occurrence only in Flemish river systems. Therefore, these optimized models should be in the future validated using data collected outside Flanders, e.g. river systems in Netherland or neighbouring regions.

References

Adamson PT, Rutherfurd ID, Peel MC, Conlan IA (2009) The hydrology of the Mekong river, 1st ed. Elsevier Inc.

Agrawal AA, Fishbein M, Halitschke R, et al. (2009) Evidence for adaptive radiation from a phylogenetic study of plant defenses. Proc Natl Acad Sci 106:18067–18072.

Al-Shami SA, Heino J, Che Salmah MR, et al. (2013) Drivers of beta diversity of macroinvertebrate communities in tropical forest streams. Freshw Biol 58:1126–1137.

Allan D (2004) Landscapes and riverscapes: the influence of land use on stream ecosystems. Annu Rev Ecol Evol Syst 35:257–284.

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43:1223–1232.

Arab A, Lek S, Lounaci A, Park YS (2009) Spatial and temporal patterns of benthic invertebrate communities in an intermittent river (North Africa). Ann Limnol - Int J Limnol 40:317–327.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. J Biogeogr 33:1677–1688.

Araújo MB, Luoto M (2007) The importance of biotic interactions for modelling species distributions under climate change. Glob Ecol Biogeogr 16:743–753.

Araújo MB, New M (2007) Ensemble forecasting of species distributions. Trends Ecol Evol 22:42–47.

Arias ME, Cochrane TA, Caruso B, et al. (2011) A landscape approach to assess impacts of hydrological changes to vegetation communities of the Tonle Sap Floodplain. Eng. Conf. Contrib. Canterbury, pp 3018–3025

Arscott DB, Tockner K, Ward J V. (2005) Lateral organization of aquatic invertebrates along the corridor of a braided floodplain river. J North Am Benthol Soc 24:934–954.

Babyak MA (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 66:411–421.

Baptista D, Buss D, Dorvillé L, Nessemian J (2001) Diversity and habitat preference of aquatic insects along the longitudinal gradient of the Macaé river basin, Rio de Janeiro, Brazil. Rev Bras Biol 61:249–258.

Barber-James HM, Gattolliat JL, Sartori M, Hubbard MD (2008) Global diversity of mayflies (Ephemeroptera, Insecta) in freshwater. Hydrobiologia 595:339–350.

Bernauer D, Jansen W (2006) Recent invasions of alien macroinvertebrates and loss of native species in the upper Rhine River, Germany. Aquat Invasions 1:55–71.

Bij de Vaate A, Jazdzewski K, Ketelaars H, et al. (2002) Geographical patterns in range extension of Ponto-Caspian macroinvertebrate species in Europe. Can J Fish Aquat Sci 59:1159–1174.

Birk S, Bonne W, Borja A, et al. (2012) Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecol Indic 18:31–41.

Boets P, Brosens D, Lock K, et al. (2016) Alien macroinvertebrates in Flanders (Belgium). Aquat Invasions 11:131–144.

Boets P, Landuyt D, Everaert G, et al. (2015) Evaluation and comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the habitat suitability for alien macroinvertebrates. Environ Model Softw 74:92–103.

Boets P, Lock K, Goethals PLM (2013) Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. Ecol Inform 17:73–81.

Boets P, Pauwels I, Lock K, Goethals P (2014) Using an integrated modelling approach for risk assessment of the "killer shrimp" Dikerogammarus villosus. River Res Appl 30:403–412.

Bogan AE (2008) Global diversity of freshwater mussels (Mollusca, Bivalvia) in freshwater. Hydrobiologia 595:139–147.

Boonsoong B, Sangpradub N, Barbour MT, Simachaya W (2010) An implementation plan for using biological indicators to improve assessment of water quality in Thailand. Environ Monit Assess 165:205–215.

Borcard D, Gillet F, Legendre P (2011) Numerical ecology with R. Springer Science & Business Media, New York

Boulton AJ, Boyero L, Covich AP, et al. (2008) Are tropical streams ecologically different from temperate streams? In: Dudgeon D (ed) Trop. Stream Ecol. Academic Press, San Diego, CA, pp 257–284

Boyero L, Ramírez A, Dudgeon D, Pearson RG (2009) Are tropical streams really different? J North Am Benthol Soc 28:397–403.

Bruns D a. (2005) Macroinvertebrate response to land cover, habitat, and water chemistry in a mining-impacted river ecosystem: A GIS watershed analysis. Aquat Sci 67:403–423.

Buffagni A, Kemp J, Erba A, et al. (2001) A Europe-wide system for assessing the quality of rivers using macroinvertebrates: The AQEM Project and its importance for southern Europe (with special emphasis on Italy). J Limnol 60:39–48.

Buisson L, Thuiller W, Casajus N, et al. (2010) Uncertainty in ensemble forecasting of species distribution. Glob Chang Biol 16:1145–1157.

Cai Y, Gong Z, Qin B (2012) Benthic macroinvertebrate community structure in Lake Taihu, China: effects of trophic status, wind-induced disturbance and habitat complexity. J Great Lakes Res 38:39–48.

Call A, Sun YX, Yu Y, et al. (2016) Genetic structure and post-glacial expansion of Cornus florida L. (Cornaceae): integrative evidence from phylogeography, population demographic history, and species distribution modeling. J Syst Evol 54:136–151.

Caraco NF, Cole JJ, Findlay SEG, et al. (2000) Dissolved oxygen declines in the Hudson River associated with the invasion of the zebra mussel (Dreissena polymorpha). Environ Sci Technol 34:1204–1210.

Caraco NF, Cole JJ, Strayer DL (2006) Top down control from the bottom: Regulation of eutrophication in a large river by benthic grazing. Limnol Oceanogr 51:664–670.

Chapman AD (2009) Numbers of living species in Australia and the World. Canberra

Chea R, Guo C, Grenouillet G, Lek S (2016) Toward an ecological understanding of a flood-pulse system lake in a tropical ecosystem: Food web structure and ecosystem health. Ecol Modell 323:1–11.

Chen L, Peng S, Yang B (2015) Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. Biol Invasions 17:2187–2198.

Chessel D (2006) The “ade4” Package.

Chowdhury GW, Zieritz A, Aldridge DC (2016) Ecosystem engineering by mussels supports biodiversity and water clarity in a heavily polluted lake in Dhaka, Bangladesh. Freshw Sci 35:188–199.

Cicchetti D V., Feinstein AR (1990) High agreement but low Kappa: II. resolving the paradoxes. J Clin Epidemiol 43:543–549.

Clarke A, Mac Nally R, Bond N, Lake PS (2008) Macroinvertebrate diversity in headwater streams: a review. Freshw Biol 53:1707–1721.

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.

Collier KJ, Lill A (2008) Spatial patterns in the composition of shallow-water macroinvertebrate communities of a large New Zealand river. New Zeal J Mar Freshw Res 42:129–141.

Cortes R, Varandas S, Teixeira A, et al. (2011) Effects of landscape metrics and land use variables on macroinvertebrate communities and habitat characteristics. Limnetica 30:347–362.

Cortes RMV, Hughes SJ, Pereira VR, Varandas SDGP (2013) Tools for bioindicator assessment in rivers: the importance of spatial scale, land use patterns and biotic integration. Ecol Indic

34:460–477.

Covich AP, Palmer MA, Crowl TA (1999) The role in of species invertebrate freshwater ecosystems - zoobenthic species influence energy flows and nutrient cycling. Bioscience 49:119–127.

Cumberlidge N, Ng PKL, Yeo DCJ (2011) Freshwater crabs of the Indo-Burma hotspot: diversity, distribution, and conservation. In: Allen DJ, Darwall WRT (eds) status Distrib. Freshw. crabs. IUCN, Gland and Cambridge, pp 102–113

Dao H, Kunpradid T, Vongsambath C, et al. (2010) Report on the 2008 biomonitoring survey of the lower Mekong River and selected tributaries, MRC Technical Paper No. 27. Vientiane, Lao PDR

David F, Boonsoong B (2014) Colonisation of leaf litter by lotic macroinvertebrates in a headwater stream of the Phachi River (western Thailand). Fundam Appl Limnol 184:109–124.

Davidson TA, Sayer CD, Perrow M, et al. (2010) The simultaneous inference of zooplanktivorous fish and macrophyte density from sub-fossil cladoceran assemblages: a multivariate regression tree approach. Freshw Biol 55:546–564.

De'ath G (2002) Multivariate regression tree: a new technique for modeling species–environment relationships. Ecology 83:1105–1117.

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

De Grave S, Cai Y, Anker A (2008) Global diversity of shrimps (Crustacea: Decapoda: Caridea) in freshwater. Hydrobiologia 595:287–293.

De Moor FC, Ivanov VD (2008) Global diversity of caddisflies (Trichoptera: Insecta) in freshwater. Hydrobiologia 595:393–407.

De Troyer N, Mereta S, Goethals P, Boets P (2016) Water quality assessment of streams and wetlands in a fast growing East African City. Water 8:123.

Descy JP, Everbecq E, Gosselain V, et al. (2003) Modelling the impact of benthic filter-feeders on the composition and biomass of river plankton. Freshw Biol 48:404–417.

Didham RK, Tylianakis JM, Gemmell NJ, et al. (2007) Interactive effects of habitat modification and species invasion on native species decline. Trends Ecol Evol 22:489–96.

Dobson M, Magana AEM, Mathooko JM, Ndegwa FK (2002) Detritivores in Kenyan highland streams: more evidence for the paucity of shredders in the tropics? Freshw Biol 47:909–919.

Dormann CF, Purschke O, Garcia Marquez JR, et al. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. Ecology 89:3371–3386.

Dray AS, Blanchet G, Borcard D, et al. (2016) Package "adespatial."

Du L-N, Yang J-X, Chen X-Y (2011) A new species of Trochotaia (Caenogastropoda: Viviparidae) from Yunnan, China. Molluscan Res 31:85–89.

Dudgeon D (2008) Tropical Stream Ecology. Academic Press, London, UK

Dudgeon D, Arthington AH, Gessner MO, et al. (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. Biol Rev Camb Philos Soc 81:163–182.

Dufrene M, Legendre P (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol Monogr 67:345–366.

Eastham J, Mpelasoka F, Ticehurst C, et al. (2008) Mekong River basin water resources assessment: impacts of climate change. CSIRO: Water for a Healthy Country National Research Flagship.

Elith J, Graham CH, Anderson RP, et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop) 29:129–151.

European Commission (2000) Directive 2000/60/EC. Establishing a framework for community action in the field of water policy.

Evangelista PH, Kumar S, Stohlgren TJ, et al. (2008) Modelling invasion for a habitat generalist and a specialist plant species. Divers Distrib 14:808–817.

Everaert G, Pauwels IS, Boets P, et al. (2013) Development and assessment of ecological models in the context of the European Water Framework Directive: Key issues for trainers in data-driven modeling approaches. Ecol Inform 17:111–116.

Everaert G, De Neve J, Boets P, et al. (2014) Comparison of the abiotic preferences of macroinvertebrates in tropical river basins. PLoS One 9:e108898.

Fajardo DRM, Seronay RA, Jumawan JC (2015) Aquatic macroinvertebrate diversity and physico-chemical characteristics of freshwater bodies in Tubay, Agusan Del Norte, Philippines. J Entomol Zool Stud 3:440–446.

Feld CK, Hering D (2007) Community structure or function: effects of environmental stress on benthic macroinvertebrates at different spatial scales. Freshw Biol 52:1380–1399.

Ferrington LC (2008) Global diversity of non-biting midges (Chironomidae; Insecta-Diptera) in freshwater. Hydrobiologia 595:447–455.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49.

Fielding AH, Haworth PF (1995) Testing the generality of bird-habitat models. Conserv Biol 9:1466–1481.

Flores MJ, Zafaralla MT (2012) Macroinvertebrate composition, diversity and richness in relation to the water quality status of Mananga River, Cebu, Philippines. Philipp Sci Lett 5:103–113.

Fochetti R, Tierno De Figueroa JM (2008) Global diversity of stoneflies (Plecoptera; Insecta) in freshwater. Hydrobiologia 595:365–377.

Forio MAE, Lock K, Radam ED, et al. (2017) Assessment and analysis of ecological quality, macroinvertebrate communities and diversity in rivers of a multifunctional tropical island. Ecol Indic 77:228–238.

Forio MAE, Van Echelpoel W, Dominguez-Granda L, et al. (2016) Analysing the effects of water quality on the occurrence of freshwater macroinvertebrate taxa among tropical river basins from different continents. AI Commun 29:665–685.

Friberg N, Skriver J, Larsen SE, et al. (2010) Stream macroinvertebrate occurrence along gradients in organic pollution and eutrophication. Freshw Biol 55:1405–1419.

Gabriels W, Lock K, De Pauw N, Goethals PLM (2010) Multimetric Macroinvertebrate Index Flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium). Limnologica 40:199–207.

Gallardo B, Aldridge DC (2013) Evaluating the combined threat of climate change and biological invasions on endangered species. Biol Conserv 160:225–233.

Goethals PLM, Dedecker AP, Gabriels W, et al. (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat Ecol 41:491–508.

Gollasch S (2006) Overview on introduced aquatic species in European navigational and adjacent waters. Helgol Mar Res 60:84–89.

Grabowski M, Bacela K, Konopacka A, Jazdzewski K (2009) Salinity-related distribution of alien amphipods in rivers provides refugia for native species. Biol Invasions 11:2107–2117.

Grenouillet G, Buisson L, Casajus N, Lek S (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography (Cop) 34:9–17.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009.

Guisan A, Zimmermann NE, Elith J, et al. (2007) What matters for predicting the occurrences of trees: techniques, data, or species characteristics? Ecol Monogr 77:615–630.

Guo C, Lek S, Ye S, et al. (2015) Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. Ecol Modell 306:67–75.

Gurevitch J, Padilla DK (2004) Are invasive species a major cause of extinctions? Trends Ecol Evol 19:470–474.

Heino J, Mykrä H (2006) Assessing physical surrogates for biodiversity: do tributary and stream type classifications reflect macroinvertebrate assemblage diversity in running waters? Biol

Conserv 129:418–426.

Heino J, Parviainen J, Paavola R, et al. (2005) Characterizing macroinvertebrate assemblage structure in relation to stream size and tributary position. Hydrobiologia 539:121–130.

Hering D, Borja A, Carstensen J, et al. (2010) The European Water Framework Directive at the age of 10: a critical review of the achievements with recommendations for the future. Sci Total Environ 408:4007–4019.

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography (Cop) 29:773–785.

Higgins SN, Vander Zanden MJ, Joppa LN, Vadeboncoeur Y (2011) The effect of dreissenid invasions on chlorophyll and the chlorophyll: total phosphorus ratio in north-temperate lakes. Can J Fish Aquat Sci 68:319–329.

Hirzel A, Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modelling. Ecol Modell 157:331–341.

Hoanh CT, Guttman H, Droogers P, Aerts J (2003) Water, climate, food, and environment in the Mekong basin in Southeast Asia - final report.

IUCN (2014) The World Conservation Union. 2014. IUCN Red List of Threatened Species, 2014.3. Summary Statistics for Globally Threatened Species. Table 1: Numbers of threatened species by major groups of organisms (1996–2014).

Jones CG, Lawton JH, Shachak M (1994) Organisms as ecosystem engineers. Oikos 69:373–386.

Kalkman VJ, Clausnitzer V, Dijkstra KDB, et al. (2008) Global diversity of dragonflies (Odonata) in freshwater. Hydrobiologia 595:351–363.

Karaouzas I, Płóciennik M (2016) Spatial scale effects on Chironomidae diversity and distribution in a Mediterranean River Basin. Hydrobiologia 767:81–93.

Kleyer M (2002) Validation of plant functional types across two contrasting landscapes. J Veg Sci 13:167–178.

Köhler F, Seddon M, Bogan AE, et al. (2012) The status and distribution of freshwater molluscs of the Indo-Burma region. In: Allen D, Smith K, Darwall W (eds) status Distrib. Freshw. Biodivers. Indo-Burma. Gland, Cambridge, pp 66–89

Królak E, Korycińska M (2008) Taxonomic composition of macroinvertebrates in the Liwiec River and its tributaries (Central and Eastern Poland) on the basis of chosen physical and chemical parameters of water and season. Polish J Environ Stud 17:39–50.

Kudthalang N, Thanee N (2010) The assessment of water quality in the upper part of the Chi basin using physicochemical variables and benthic macroinvertebrates. Suranaree J Sci Technol 17:165–176.

Kummu M, Lu XX, Rasphone A, et al. (2008) Riverbank changes along the Mekong River: remote sensing detection in the Vientiane-Nong Khai area. Quat Int 186:100–112.

Legendre P, De Cáceres M (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecol Lett 16:951–963.

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.

Legendre P, Legendre L (2012) Numerical ecology-developments in environmental modelling, 3rd ed. Elsevier Science BV, Amsterdam

Legendre P, Salvat B (2015) Thirty-year recovery of mollusc communities after nuclear experimentations on Fangataufa atoll (Tuamotu, French Polynesia). Proc. R. Soc. London B Biol. Sci. 282:

Lek S, Delacoste M, Baran P, et al. (1996) Application of neural networks to modelling nonlinear relationships in ecology. Ecol Modell 90:39–52.

Lele SR, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Lett 10:551–563.

Lele SR, Nadeem K, Schmuland B (2010) Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. J Am Stat Assoc 105:1617–1625.

Lencioni V, Maiolini B, Marziali L, et al. (2007) Macroinvertebrate assemblages in glacial stream systems: a comparison of linear multivariate methods with artificial neural networks. Ecol Modell 203:119–131.

Leung ASL, Dudgeon D (2011) Scales of spatiotemporal variability in macroinvertebrate abundance and diversity in monsoonal streams: detecting environmental change. Freshw Biol 56:1193–1208.

Leuven RSEW, van der Velde G, Baijens I, et al. (2009) The river Rhine: a global highway for dispersal of aquatic invasive species. Biol Invasions 11:1989–2008.

Luoto M, Heikkinen RK, Pöyry J, Saarinen K (2006) Determinants of the biogeographical distribution of butterflies in boreal regions. J Biogeogr 33:1764–1778.

Magbanua FS, Yvette N, Mendoza B, et al. (2015) Water physicochemistry and benthic macroinvertebrate communities in a tropical reservoir: the role of water level fluctuations and water depth. Limnologica 55:13–20.

Manel S, Dias JM, Buckton ST, Ormerod SJ (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J Appl Ecol 36:734–747.

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence-absence models in ecology; the need to count for prevalence. J Appl Ecol 38:921–931.

Martin P, Martinez-Ansemil E, Pinder A, et al. (2008) Global diversity of oligochaetous clitellates (“Oligochaeta”; Clitellata) in freshwater. Hydrobiologia 595:117–127.

Mastrorillo S, Lek S, Dauba F, Belaud A (1997) The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshw Biol 38:237–246.

May RM (1988) How many species are there on Earth? Science (80-) 241:1441–1449.

Mazão GR, Bispo P da C (2016) The influence of physical instream spatial variability on Chironomidae (Diptera) assemblages in Neotropical streams. Limnologica 60:1–5.

McCluskey A, Lalkhen AG (2007) Statistics II: central tendency and spread of data. Contin Educ Anaesthesia, Crit Care Pain 7:127–130.

McKenna DD, Farrell BD (2006) Tropical forests are both evolutionary cradles and museums of leaf beetle diversity. Proc Natl Acad Sci 103:10947–10951.

McPherson JM, Jetz W, Rogers DJ (2004) The effects of species’ range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J Appl Ecol 41:811–823.

Md Rawi C, Al-Shami SA, Madrus MR, Ahmad AH (2013) Biological and ecological diversity of aquatic macroinvertebrates in response to hydrological and physicochemical parameters in tropical forest streams of Gunung Tebu, Malaysia: implications for ecohydrological assessment. Ecohydrology 7:496–507.

Meier ES, Kienast F, Pearman PB, et al. (2010) Biotic and abiotic variables show little redundancy in explaining tree species distributions. Ecography (Cop) 33:1038–1048.

Mereta ST, Boets P, Ambelu Bayih A, et al. (2012) Analysis of environmental factors determining the abundance and diversity of macroinvertebrate taxa in natural wetlands of Southwest Ethiopia. Ecol Inform 7:52–61.

Muñoz I, López-Doval J, Ricart M, et al. (2009) Bridging levels of pharmaceuticals in river water with biological community structure in the Llobregat River basin (Northern Spain). Environ Toxicol Chem 28:2706–2714.

Naeem S (1998) Species redundancy and ecosystem reliability. Conserv Biol 12:39–45.

Nhan DK, Phong LT, Verdegem MJC, et al. (2007) Integrated freshwater aquaculture, crop and livestock production in the Mekong delta, Vietnam: determinants and the role of the pond. Agric Syst 94:445–458.

Nicola GG, Almodóvar A, Elvira B (2010) Effects of environmental factors and predation on benthic communities in headwater streams. Aquat Sci 72:419–429.

Nieser N, Chen PP, Yang CM (2005) A new subgenus and six new species of Nepomorpha

(Insecta: Heteroptera) from Yunnan, China. Raffles Bull Zool 53:189–209.

Nunes AL, Tricarico E, Panov VE, et al. (2015) Pathways and gateways of freshwater invasions in Europe. Aquat Invasions 10:359–370.

Özesmi U, Mitsch WJ (1997) A spatial habitat model for the marsh-breeding red-winged blackbird (Agelaius phoeniceus L.) in coastal Lake Erie wetlands. Ecol Modell 101:139–152.

Palmer M, Covich A, Finlay B, et al. (1997) Biodiversity and ecosystem processes in freshwater sediments. Ambio 26:571–577.

Pan B, Wang Z, Li Z, et al. (2013) An exploratory analysis of benthic macroinvertebrates as indicators of the ecological status of the Upper Yellow and Yangtze Rivers. J Geogr Sci 23:871–882.

Panov VE, Alexandrov B, Arbaciauskas K, et al. (2009) Assessing the risks of aquatic species invasions via European inland waterways: from concepts to environmental indicators. Integr Environ Assess Manag 5:110–126.

Park YS, Cereghino R, Compin A, Lek S (2003) Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol Modell 160:265–280.

Parnrong S, Buathong M, Sites RW (2002) New records of Behningiidae, Potamanthidae, and Prosopistomatidae (Ephemeroptera) from Thailand. ScienceAsia 28:407–409.

Parravicini V, Azzurro E, Kulbicki M, Belmaker J (2015) Niche shift can impair the ability to predict invasion risk in the marine realm: An illustration using Mediterranean fish invaders. Ecol Lett 18:246–253.

Pearson RG (2014) Dynamics of invertebrate diversity in a tropical stream. Diversity 6:771–791.

Pearson RG, Dawson TP, Berry PM, Harrison PA (2002) SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecol Modell 154:289–300.

Peel MC, Finlayson BL, McMahon T a. (2007) Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci 11:1633–1644.

Pérez-Quintero JC (2011) Distribution patterns of freshwater molluscs along environmental gradients in the southern Guadiana River basin (SW Iberian Peninsula). Hydrobiologia 678:65–76.

Phaphong A, Sangpradub N (2012) Development of a benthic macroinvertebrate biotic index to evaluate wetland health in Northeastern Thailand. African J Agric Res 7:6320–6328.

Poelmans L, Van Rompaey A (2009) Detecting and modelling spatial patterns of urban sprawl in highly fragmented areas: a case study in the Flanders-Brussels region. Landsc Urban Plan 93:10–19.

Pollard P, Huxham M (1998) The European water framework directive: a new era in the management of aquatic ecosystem health? Aquat Conserv Freshw Ecosyst 8:773–792.

Ponciano JM, Taper ML, Dennis B, Lele S (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. Ecology 90:356–362.

Poulsen AF, Hortle KG, Chan S, et al. (2004) Distribution and Ecology of Some Important Riverine Fish Species of the Mekong River Basin. 1–116.

Qi D, Chao Y, Guo S, et al. (2012) Convergent, parallel and correlated evolution of trophic morphologies in the subfamily schizothoracinae from the qinghai-tibetan plateau. PLoS One 7:1–10.

R Core Team (2013) R: a language and environment for statistical computing.

Rahel FJ (2002) Homogenization of freshwater faunas. Annu Rev Ecol Syst 33:291–315.

Rahel FJ (2007) Biogeographic barriers, connectivity and homogenization of freshwater faunas: tt's a small world after all. Freshw Biol 52:696–710.

Randin CF, Dirnböck T, Dullinger S, et al. (2006) Are niche-based species distribution models transferable in space? J Biogeogr 33:1689–1703.

Rao CR (1995) A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. Qüestiió 19:23–63.

Regan TJ, McCarthy MA, Baxter PWJ, et al. (2006) Optimal eradication: when to stop looking for an invasive plant. Ecol Lett 9:759–766.

Resh VH (2007) Multinational, freshwater biomonitoring programs in the developing world: lessons learned from African and Southeast Asian river surveys. Environ Manage 39:737–748.

Roberts W (2013) Package "labdsv."

Rodrigues Capítulo A, Tangorra M, Ocón C (2001) Use of benthic macroinvertebrates to assess the biological status of Pampean streams in Argentina. Aquat Ecol 35:109–119.

Roura-Pascual N, Brotons L, Peterson AT, Thuiller W (2009) Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biol Invasions 11:1017–1031.

Sala OE, Chapin III FS, Armesto JJ, et al. (2000) Global biodiversity scenarios for the year 2100. Science (80-) 287:1770–1774.

Salmah MRC, Al-Shami SA, Abu Hassan A, et al. (2014) Distribution of detritivores in tropical forest streams of peninsular Malaysia: role of temperature, canopy cover and altitude variability. Int J Biometeorol 58:679–690.

Sangpradub N, Hanjavanit C, Boonsoong B (2002) New records of Heptageniid mayflies Asionurus and Thalerosphyrus (Ephemeroptera: Heptageniidae) from Northeastern Thailand. ScienceAsia 28:411–416.

Sarkkula J, Kiirikki M, Koponen J, Kummu M (2003) Ecosystem processes of the Tonle Sap Lake. Phnom Penh, Cambodia

Schröder W, Pesch R, Schmidt G (2007) Statistical classification of terrestrial and marine ecosystems for environmental planning. Landsc Online 2:1–22.

Sedell JR, Rchey JE, Swanson FJ (1989) The river continuum concept: a basis for the expected ecosystem behavior of very large rivers? Can Spec Publ Fish Aquat Sci 106:49–55.

Segurado P, Araujo M (2004) An evaluation of methods for modelling species distributions. J Biogeogr 31:1555–1568.

Shao ML, Xie ZC, Han XQ, et al. (2008) Macroinvertebrate community structure in Three-Gorges Reservoir, China. Int Rev Hydrobiol 93:175–187.

Sinco AL, Sendaydiego JP, Saab LL, et al. (2014) Riverine biota as indicators of water quality in tropical Cagayan de Oro River, Philippines. Adv Environ Sci - Int J Bioflux Soc 6:157–167.

Sodhi NS, Koh LP, Brook BW, Ng PKL (2004) Southeast Asian biodiversity: an impending disaster. Trends Ecol Evol 19:654–60.

Sor R, Boets P, Chea R, et al. (2017a) Spatial organization of macroinvertebrate assemblages in the Lower Mekong Basin. Limnologica 64:20–30.

Sor R, Park Y-S, Boets P, et al. (2017b) Effects of species prevalence on the performance of predictive models. Ecol Modell 354:11–19.

Sousa R, Novais A, Costa R, Strayer DL (2014) Invasive bivalves in fresh waters: impacts from individuals to ecosystems and possible control strategies. Hydrobiologia 735:233–251.

Stockwell DR, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Modell 148:1–13.

Strong EE, Gargominy O, Ponder WF, Bouchet P (2008) Global diversity of gastropods (Gastropoda; Mollusca) in freshwater. Hydrobiologia 595:149–166.

Suhaila A, Che Salmah MR (2014) Ecology of Ephemeroptera, Plecoptera and Trichoptera (Insecta) in rivers of the Gunung Jerai forest reserve: diversity and distribution of functional feeding groups. Trop Life Sci Res 25:61–73.

Svenning JC, Normand S, Kageyama M (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. J Ecol 96:1117–1127.

Szöcs E, Coring E, Bäthe J, Schäfer RB (2014) Effects of anthropogenic salinization on biological traits and community composition of stream macroinvertebrates. Sci Total Environ 468–469:943–9.

Tampus AD, Tobias EG, Amparado RF, et al. (2012) Water quality assessment using macroinvertebrates and physico-chemical parameters in the riverine system of Iligan City, Philippines. AES Bioflux 4:59–68.

Therneau TM, Atkinson EJ (1997) An introduction to recursive partitioning using the rpart routines. Technical report no. 61. Rochester, Minnesota

Thorp JH, Delong M (1994) The riverine productivity model: an heuristic view of carbon sources and organic processing in large river ecosystems. Oikos 2:305–308.

Thuiller W, Richardson DM, Py Ek P, et al. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Glob Chang Biol 11:2234–2250.

Van Echelpoel W, Boets P, Landuyt D, et al. (2015) Species distribution models for sustainable ecosystem management. In: Park Y, Lek S, Baehr C, Jorgensen S (eds) 19th Glob. Bienn. Conf. Int. Elsevier Science BV, Toulouse, France, pp 115–134

van Griensven A, Vandenberghe V (2006) Monitoring in Rural Areas. In: Quevauviller P, Thomas O, Van der Beken A (eds) Wastewater Qual. Monit. Treat. John Wiley & Sons, Inc., West Sussex, England, pp 145–160

Vannote RL, Minshall GW, Cummins KW, et al. (1980) The River Continuum Concept. Can J Fish Aquat Sci 37:130–137.

Vaughn CC, Spooner DE (2006) Unionid mussels influence macroinvertebrate assemblage structure in streams. J North Am Benthol Soc 25:691–700.

Verdonschot PFM, Nijboer RC (2004) Testing the European stream typology of the Water Framework Directive for macroinvertebrates. Hydrobiologia 516:35–54.

Vicente J, Randin CF, Gonçalves J, et al. (2011) Where will conflicts between alien and rare species occur after climate and land-use change? a test with a novel combined modelling approach. Biol Invasions 13:1209–1227.

Wang B, Liu D, Liu S, et al. (2012) Impacts of urbanization on stream habitats and macroinvertebrate communities in the tributaries of Qiangtang River, China. Hydrobiologia 680:39–51.

Wetzel RG (2001) Limnology: lake and river ecosystems, 3rd ed. CA: Academic Press, San Diego

Wilby A, Lan LP, Heong KL, et al. (2006) Arthropod diversity and community structure in relation to land use in the Mekong Delta, Vietnam. Ecosystems 9:538–549.

Wu X, He D, Yang G, et al. (2014) Seasonal variability of water quality and metazooplankton community structure in Xiaowan Reservoir of the upper Mekong River. J Limnol 73:167–176.

Yeo DCJ, Ng PKL, Cumberlidge N, et al. (2008) Global diversity of crabs (Crustacea: Decapoda: Brachyura) in freshwater. Hydrobiologia 595:275–286.

Zilli FL, Marchese MR (2011) Patterns in macroinvertebrate assemblages at different spatial scales: implications of hydrological connectivity in a large floodplain river. Hydrobiologia 663:245–257.

Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577.

Appendices

Appendix T1

List of sampling sites, corresponding codes used in this study (Site code) and in the Mekong River Commission report (MRC site code), and the channel and the year sampled. * indicates the site which was sampled further away from its original coordinates and which was considered as a different site in our study.

Site	MRC	Coordinates (UTM)			Year sampled				
code	site code	Easting	Northing	Channel	2004	2005	2006	2007	2008
CB1	CKL	48P 503327	1246641	Main			x		
CM1	CKT	48P 610951	1393569	Main	x		x		
CM2	CMR	48P 607964	1537129	Main		x			
CM3	CMR.1*	48P 618663	1504098	Main					x
CM4	CNL	48P 528321	1250852	Main			x		
CT1	CKM	48P 615596	1500691	Tributary		x	x	x	
CT2	CKM.1*	48P 606331	1539069	Tributary					x
CT3	CPP	48P 492492	1279903	Tributary	x		x		
CT4	CPS	48P 381258	1382944	Tributary	x				
CT5	CPT	48P 613899	1374811	Tributary			x		
CT6	CSJ	48P 621005	1499145	Tributary		x	x	x	
CT7	CSK	48P 348375	1465699	Tributary			x		
CT8	CSN	48P 490998	1401845	Tributary			x		
CT9	CSP	48P 716971	1490691	Tributary	x	x	x	x	
CT10	CSU	48P 764687	1526041	Tributary		x			
CT11	CSS	48P 696445	1545480	Tributary	x				
CT12	CTU	48P 477884	1309367	Tributary	x				
LM1	LDN	48P 596621	1650516	Main				x	
LM2	LMH	47Q 723733	2383320	Main		x			
LM3	LMX	47Q 670860	2311778	Main		x			
LM4	LPB	48Q 201739	2203028	Main	x				
LM5	LPS	48P 587623	1671756	Main	x				
LM6	LVT	48Q 239871	1988731	Main	x				
LM7	LVT.1*	48Q 229378	1990015	Main					x
LT1	LBF	48Q 498437	1888075	Tributary				x	
LT2	LSD	48P 586345	1673985	Tributary				x	
LT3	LBH	48Q 540315	1779816	Tributary				x	
LT4	LKD	48Q 398871	2023713	Tributary	x			x	
LT5	LKL	48P 673642	1622904	Tributary		x		x	
LT6	LKU	48P 701679	1653515	Tributary		x		x	
LT7	LNG	48Q 240744	2050118	Tributary	x			x	
LT8	LNK	48Q 203428	2200953	Tributary		x			
LT9	LNM	48Q 280667	2088210	Tributary				x	
LT10	LNT	48Q 208083	2016581	Tributary				x	
LT11	LNO	48Q 212495	2222855	Tributary	x				
LT12	LOU	48Q 219345	2229380	Tributary		x			
TM1	TCS	47Q 614718	2240109	Main					x
TM2	TKC	48P 552099	1694552	Main					x
TM3	TMC	47Q 655974	2231281	Main		x			
TM4	TNP	48Q 450496	1874332	Main					x
TM5	TSM	48Q 444135	1951422	Main				x	
TT1	TCH	48P 407724	1745362	Tributary	x				
TT2	TKO	47Q 576165	2205993	Tributary	x	x			
TT3	TMI	47Q 640355	2213637	Tributary		x			
TT4	TMU	48P 553283	1692193	Tributary	x				
TT5	TMM	48P 552854	1692378	Tributary				x	
TT6	TNK	48Q 450473	1874626	Tributary				x	
TT7	TSK	48Q 438501	1946480	Tributary	x			x	

Site	MRC	Coordinates (UTM)			Year sampled				
code	site code	Easting	Northing	Channel	2004	2005	2006	2007	2008
TT8	TUN	48P 494860	1685056	Tributary					x
VB1	VCD	48P 515263	1187502	Main	x		x		
VB2	VDP	48P 514690	1188035	Main					x
VB3	VCT	48P 588365	1110673	Main			x		
VB4	VKB	48P 509482	1210872	Main					x
VB5	VLX	48P 551878	1143546	Main			x		
VM1	VCL	48P 563807	1153868	Main			x		
VM2	VTC	48P 528931	1194535	Main	x				
VM3	VTT	48P 528951	1194447	Main					x
VM4	VTP	48P 519830	1205766	Main					x
VM5	VTR	48P 603976	1135759	Main			x		
VM6	VVL	48P 603698	1134514	Main					x
VT1	VSP	48P 802270	1426825	Tributary	x				
VT2	VSR	48P 817329	1396950	Tributary			x		
VT3	VSS	49P 180575	1587838	Tributary	x		x		

Appendix T2

List of taxa recorded in the present study and their occurrences. The numbers in the parentheses tell the number of taxa reported for each family/order/clade. The families/orders/clades that are not followed by any number mean that they are represented by only one observed taxon.

Taxonomic group	Order/Clade	Family	Species	Species occurrence
Annelid	Aphanoneura	Aeolosomidae	*Aeolosoma bengalense*	3
Annelid	Haplotaxida (15)	Naididae (15)	*Aulodrilus prothecatus*	5
Annelid	-	-	*Branchiodrilus semperi*	2
Annelid	-	-	*Branchiura sowerbyi*	29
Annelid	-	-	*Chaetogaster langi*	8
Annelid	-	-	*Chaetogaster limnaei limnaei*	4
Annelid	-	-	*Chaetogaster* sp.	4
Annelid	-	-	*Dero denticulata*	1
Annelid	-	-	*Dero pectinata*	5
Annelid	-	-	*Dero* sp.	6
Annelid	-	-	*Dero* sp.1	11
Annelid	-	-	*Dero* sp.2	10
Annelid	-	-	*Dero* sp.3	2
Annelid	-	-	*Limnodrilus hoffmeisteri*	28
Annelid	-	-	Naididae sp.	20
Annelid	-	-	*Pristina* sp.	2
Annelid	Lumbriculida	Lumbriculidae	Lumbriculidae sp.	4
Annelid	Phyllodocida (4)	Nephtyidae	*Nephtys polybranchia*	2
Annelid	-	Nereididae (3)	*Namalycastis abiuma*	6
Annelid	-	-	*Namalycastis longicirris*	10
Annelid	-	-	*Neanthes caudata*	1
Annelid	Scolecida (3)	Opheliidae	*Polyophthalmus pictus*	1
Annelid	-	Orbiniidae (2)	*Orbinia johnsoni*	3
Annelid	-	-	*Scoloplos* sp.	1
Annelid	Spionida (2)	Spionidae (2)	*Polydora* sp.	5
Annelid	-	-	*Prionospio* sp.	1
Annelid	Unknown (4)	Unknown (4)	Oligochaeta sp.	12
Annelid	-	-	Polychaeta sp.	1
Annelid	-	-	Polychaeta sp.1	4
Annelid	-	-	Polychaeta sp.2	1
Crustacean	Amphipoda (10)	Corophiidae (7)	*Corophium intermedium*	1
Crustacean	-	-	*Corophium minutum*	6
Crustacean	-	-	*Corophium* sp.	5
Crustacean	-	-	*Grandidierella lignorum*	8
Crustacean	-	-	*Grandidierella vietnamica*	11
Crustacean	-	-	*Kamaka* sp.	5
Crustacean	-	-	*Monocorophium* sp.	9
Crustacean	-	Melitidae (2)	*Melita* sp.	11
Crustacean	-	-	*Melita vietnamica*	2
Crustacean	-	Oedicerotidae	*Perioculodes* sp.	3
Crustacean	Cumacea	Diastylidae	*Diastylis rathkei*	1
Crustacean	Decapoda (18)	Alpheidae	*Alpheus bisincisus*	1
Crustacean	-	Atyidae (3)	*Caridina nilotica*	1
Crustacean	-	-	*Caridina* sp.	1
Crustacean	-	-	*Caridina verrata verrata*	2
Crustacean	-	Palaemonidae (10)	*Macrobrachium dienbienphuensis*	1
Crustacean	-	-	*Macrobrachium equidens*	4
Crustacean	-	-	*Macrobrachium lanchesteri*	5
Crustacean	-	-	*Macrobrachium mekongene*	2
Crustacean	-	-	*Macrobrachium pilimanus*	5
Crustacean	-	-	*Macrobrachium rosenbergii*	2
Crustacean	-	-	*Macrobrachium secamense*	1
Crustacean	-	-	*Macrobrachium* sp.	5
Crustacean	-	-	*Palaemon curvirostris*	4

Taxonomic group	Order/Clade	Family	Species	Species occurrence
Crustacean	-	-	Palaemonidae larva	2
Crustacean	-	Parathelphusidae (2)	Parathelphusidae sp.	1
Crustacean	-	-	*Somanniathelphusa germaini*	4
Crustacean	-	Potamidae	*Ranguna cochinchinensis*	1
Crustacean	-	Unknown	Decapoda larva	9
Crustacean	Isopoda (3)	Anthuridae (2)	*Cyathura carinata*	6
Crustacean	-	-	*Cyathura truncata*	8
Crustacean	-	Corallanidae	*Tachaea chinensis*	1
Crustacean	Tanaidacea (7)	Apseudidae (2)	*Apseudes vietnamensis*	1
Crustacean	-	-	Apseudidae sp.	1
Crustacean	-	Gammaridae	*Gammarus* sp.	4
Crustacean	-	Haustoriidae (2)	*Eohaustorius* sp.	4
Crustacean	-	-	*Eohaustorius tandeensis*	5
Crustacean	-	Hyalidae (2)	*Hyale hawaiensis*	5
Crustacean	-	-	*Hyale* sp.	8
Insect	Coleoptera (9)	Amphizoidae	*Amphizoa* sp.	1
Insect	-	Dytiscidae (2)	*Hyphydrus* sp.	1
Insect	-	-	*Neptosternus* sp.	2
Insect	-	Elmidae (4)	*Cloeon* sp.	9
Insect	-	-	Elmidae sp.	9
Insect	-	-	*Heterlimnius* sp.	4
Insect	-	-	*Stenelmis* sp.	2
Insect	-	Haliplidae (2)	Haliplidae sp.	2
Insect	-	-	*Haliplus* sp.	1
Insect	Diptera (36)	Ceratopogonidae (3)	*Bezzia* sp.	18
Insect	-	-	*Culicoides* sp.	30
Insect	-	-	*Dasyhelea* sp.	3
Insect	-	Chaoboridae	*Chaoborus* sp.	6
Insect	-	Chironomidae (23)	*Ablabesmyia* sp.	46
Insect	-	-	Chironomidae sp.	25
Insect	-	-	*Chironomus attenuatus*	1
Insect	-	-	*Chironomus* sp.	33
Insect	-	-	*Chironomus* sp.1	1
Insect	-	-	*Clinotanypus* sp.	3
Insect	-	-	*Clypeocaetis* sp.	1
Insect	-	-	*Cricotopus* sp.	11
Insect	-	-	*Cryptochironomus* sp.	25
Insect	-	-	Diamesinae sp.	2
Insect	-	-	*Einfeldia* sp.	1
Insect	-	-	*Glyptotendipes* sp.	1
Insect	-	-	*Goeldichironomus* sp.	27
Insect	-	-	Orthocladiinae sp.	1
Insect	-	-	*Parachironomus* sp.	1
Insect	-	-	*Polypedilum* sp.	44
Insect	-	-	*Procladius* sp.	1
Insect	-	-	*Pseudochironomus* sp.	3
Insect	-	-	*Pseudodiamesa* sp.	4
Insect	-	-	*Sergentia* sp.	9
Insect	-	-	*Smittia* sp.	4
Insect	-	-	*Tanypus* sp.	1
Insect	-	-	*Tanytarsus* sp.	2
Insect	-	Dolichopodidae	*Hydrophorus* sp.	1
Insect	-	Limoniidae	*Limnophila* sp.	2
Insect	-	Tabanidae (2)	*Chrysops* sp.	2
Insect	-	-	*Tabanus* sp.	1
Insect	-	Tipulidae (5)	*Antocha* sp.	3
Insect	-	-	*Eriocera* sp.	15
Insect	-	-	*Pedicia* sp.	1
Insect	-	-	*Tipula* sp.	1

Taxonomic group	Order/Clade	Family	Species	Species occurrence
Insect	-	-	Tipulidae sp.	1
Insect	Ephemeroptera (32)	Baetidae (6)	*Acentrella* sp.	1
Insect	-	-	*Baetiella* sp.	1
Insect	-	-	*Baetis* sp.	11
Insect	-	-	*Centroptilum* sp.	3
Insect	-	-	*Heterocloeon* sp.	3
Insect	-	-	*Procloeon* sp.	1
Insect	-	Behningiidae	Behningiidae sp.	1
Insect	-	Caenidae (4)	*Caenis* sp.	21
Insect	-	-	*Caenoculis* sp.	3
Insect	-	-	*Caenodes* sp.	6
Insect	-	-	*Cercobrachys* sp.	2
Insect	-	Ephemeridae (4)	*Afromera* sp.	1
Insect	-	-	*Eatonigenia* sp.	2
Insect	-	-	*Ephemera* sp.	14
Insect	-	-	*Hexagenia* sp.	1
Insect	-	Heptageniidae (6)	*Cinygmina* sp.	3
Insect	-	-	*Cladopelma* sp.	4
Insect	-	-	*Epeorus* sp.	1
Insect	-	-	*Heptagenia* sp.	1
Insect	-	-	Heptageniidae sp.	1
Insect	-	-	*Thalero sphyrus* sp.	1
Insect	-	Leptoplebiidae (3)	*Choropterpes* sp.	5
Insect	-	-	*Leptophlebia* sp.	2
Insect	-	-	*Traverella* sp.	2
Insect	-	Oligoneuriidae (2)	*Chromarcys* sp.	1
Insect	-	-	*Pentagenia* sp.	9
Insect	-	Palingeniidae (2)	*Anagenesia* sp.	5
Insect	-	-	Palingeniidae sp.	1
Insect	-	Polymitarcyidae (2)	*Ephoron* sp.	2
Insect	-	-	*Povilla* sp.	4
Insect	-	Potamanthidae	*Potamanthus* sp.	4
Insect	-	Prosopistomatidae	*Prosopistoma* sp.	1
Insect	Hemiptera (8)	Belostomatidae	*Diplonychus rusticus*	5
Insect	-	Corixidae (4)	*Micronecta* sp.	4
Insect	-	-	*Corixa* sp.	1
Insect	-	-	*Microtendipes* sp.	1
Insect	-	-	*Sigara* sp.	4
Insect	-	Delphacidae	*Megamelus* sp.	2
Insect	-	Gerridae	Gerridae sp.	1
Insect	-	Naucoridae	*Naucoris* sp.	8
Insect	Lepidoptera	Pyralidae	Pyralidae sp.	1
Insect	Odonata (22)	Aeshnidae	*Aeshna* sp.	1
Insect	-	Calopterygidae (2)	*Agrion* sp.	2
Insect	-	-	*Calopteryx* sp.	1
Insect	-	Corduliidae (3)	*Cordulia* sp.	1
Insect	-	-	*Epitheca* sp.	1
Insect	-	-	*Macromia* sp.	5
Insect	-	Gomphidae (14)	*Amphylla williamsoni*	2
Insect	-	-	*Aphylla* sp.	11
Insect	-	-	*Arigomphus* sp.	5
Insect	-	-	*Burmagomphus* sp.	1
Insect	-	-	*Dromogomphus* sp.	23
Insect	-	-	*Gastrogomphus* sp.	1
Insect	-	-	Gomphidae sp.	8
Insect	-	-	*Gomphus* sp.	1
Insect	-	-	*Labrogomphus* sp.	1
Insect	-	-	*Megalogomphus* sp.	1
Insect	-	-	*Octogomphus* sp.	5

Taxonomic group	Order/Clade	Family	Species	Species occurrence
Insect	-	-	*Ophiogomphus* sp.	2
Insect	-	-	*Orientogomphus* sp.	1
Insect	-	-	*Progomphus* sp.	6
Insect	-	Libellulidae	*Libellula* sp.	5
Insect	-	Petaluridae	*Tachopteryx* sp.	1
Insect	Plecoptera (3)	Perlidae (3)	*Etrocorema* sp.	1
Insect	-	-	*Perla* sp.	3
Insect	-	-	*Phanoperla* sp.	1
Insect	Trichoptera (21)	Dipseudopsidae	*Dipseudopsis* sp.	8
Insect	-	Ecnomidae	*Economus* sp.	9
Insect	-	Glossosomatidae	*Glososoma* sp.	1
Insect	-	Goeridae	*Goera* sp.	1
Insect	-	Hydropsychidae (3)	*Cheumaatopsyche* sp.	1
Insect	-	-	*Hydropsyche* sp.	5
Insect	-	-	*Macronema* sp.	2
Insect	-	Hydroptilidae (3)	*Agraylea* sp.	1
Insect	-	-	*Orthotrichia* sp.	1
Insect	-	-	*Oxyethira* sp.	1
Insect	-	Leptoceridae (4)	*Leptocerus* sp.	2
Insect	-	-	Leptoceridae sp.	2
Insect	-	-	*Oecetis* sp.	1
Insect	-	-	*Nectopsyche* sp.	5
Insect	-	Limnephilidae (2)	*Farula* sp.	1
Insect	-	-	*Limnephilus* sp.	1
Insect	-	Molannidae	*Molanna* sp.	1
Insect	-	Philopotamidae (2)	*Chimarra* sp.	1
Insect	-	-	Philopotamidae sp.	20
Insect	-	Psychomyiidae	Psychomyiidae sp.	9
Insect	-	Rhyacophilidae	*Rhyacophila* sp.	2
Mollusc	Arcoida (2)	Arcidae (2)	*Scaphula pinna*	5
Mollusc	-	-	*Scaphula* sp.	3
Mollusc	Mytiloida (3)	Mytilidae (3)	*Limnoperna siamensis*	17
Mollusc	-	-	*Limnoperna* sp.	10
Mollusc	-	-	*Sinomytilus harmandi*	12
Mollusc	Unionida (18)	Unionidae (18)	*Ensidens ingallsianus ingallsianus*	6
Mollusc	-	-	*Hyriopsis Hyriopsis bialatus*	6
Mollusc	-	-	*Hyriopsis Limnoscapha desowitzi*	1
Mollusc	-	-	*Indonaia pilata*	6
Mollusc	-	-	*Physunio cambodiensis*	3
Mollusc	-	-	*Physunio micropterus*	3
Mollusc	-	-	*Pilsbryoconcha exilis compressa*	4
Mollusc	-	-	*Pilsbryoconcha exilis exilis*	4
Mollusc	-	-	*Pilsbryoconcha lemeslei*	2
Mollusc	-	-	*Pseudodon cambodjensis cambodjensis*	2
Mollusc	-	-	*Pseudodon inoscularis cumingi*	1
Mollusc	-	-	*Pseudodon vondembuschianus ellipticus*	1
Mollusc	-	-	*Scabies scobinata*	1
Mollusc	-	-	*Scabies* sp.	2
Mollusc	-	-	*Trapezoideus exolescens comptus*	2
Mollusc	-	-	*Uniandra contradens ascia*	5
Mollusc	-	-	*Uniandra contradens tumidula*	1
Mollusc	-	-	Unionida sp.	1
Mollusc	Veneroida (15)	Corbiculidae (14)	*Corbicula arata*	1
Mollusc	-	-	*Corbicula baudoni*	12
Mollusc	-	-	*Corbicula blandiana*	16
Mollusc	-	-	*Corbicula bocourti*	6
Mollusc	-	-	*Corbicula castanea*	4

Taxonomic group	Order/Clade	Family	Species	Species occurrence
Mollusc	-	-	*Corbicula cyreniformis*	11
Mollusc	-	-	*Corbicula fluminea*	1
Mollusc	-	-	*Corbicula gustaviana*	1
Mollusc	-	-	*Corbicula lamarckiana*	20
Mollusc	-	-	*Corbicula* larva	10
Mollusc	-	-	*Corbicula leviuscula*	14
Mollusc	-	-	*Corbicula moreletiana*	14
Mollusc	-	-	*Corbicula* sp.	11
Mollusc	-	-	*Corbicula tenuis*	42
Mollusc	-	Pisidiidae	*Afropisidium clarkeanum*	9
Mollusc	Caenogastropoda (50)	Ampullariidae (3)	*Pila ampullacea*	2
Mollusc	-	-	*Pila polita*	1
Mollusc	-	-	*Pila scutata*	1
Mollusc	-	Assimineidae (2)	*Cyclotropis bollingi*	3
Mollusc	-	-	*Cyclotropis* sp.	2
Mollusc	-	Bithyniidae (3)	*Bithynia siamensis*	5
Mollusc	-	-	*Bithynia* sp.	12
Mollusc	-	-	*Wattebledia siamensis*	1
Mollusc	-	Buccinoidae (3)	*Clea scalarina*	1
Mollusc	-	-	*Clea helena*	1
Mollusc	-	-	*Clea* sp.1	5
Mollusc	-	Cochliopidae	*Cochliopa riograndensis*	2
Mollusc	-	Hydrobiidae	*Kareliania* sp.	3
Mollusc	-	Pachychilidae (2)	*Adamietta housei*	1
Mollusc	-	-	*Brotia* sp.	2
Mollusc	-	Pomatiopsidae (8)	*Hubendickia crooki*	4
Mollusc	-	-	*Hubendickia* sp.	8
Mollusc	-	-	*Hydrorissoia* sp.	1
Mollusc	-	-	*Lacunopsis* sp.	1
Mollusc	-	-	*Pachydrobia brevis*	2
Mollusc	-	-	*Pachydrobia* sp.	10
Mollusc	-	-	*Pachydrobiella* sp.	1
Mollusc	-	-	*Paraprososthenia* sp.	2
Mollusc	-	Stenothyridae (9)	*Stenothyra annandalei*	4
Mollusc	-	-	*Stenothyra glabrata*	8
Mollusc	-	-	*Stenothyra jiraponi*	2
Mollusc	-	-	*Stenothyra koratensis holosculpta*	18
Mollusc	-	-	*Stenothyra koratensis koratensis*	4
Mollusc	-	-	*Stenothyra labiata*	3
Mollusc	-	-	*Stenothyra mcmulleni*	12
Mollusc	-	-	*Stenothyra moussoni*	3
Mollusc	-	-	*Stenothyra* sp.	4
Mollusc	-	Thiaridae (6)	*Melanoides* sp.	1
Mollusc	-	-	*Melanoides tuberculata*	12
Mollusc	-	-	*Neoradina prasongi*	2
Mollusc	-	-	*Sermyla tornatella*	16
Mollusc	-	-	*Tarebia granifera*	3
Mollusc	-	-	*Thiara scabra*	2
Mollusc	-	Viviparidae (12)	*Angulyagra polyzonata*	4
Mollusc	-	-	*Angulyagra* sp.	10
Mollusc	-	-	*Anulotaia* sp.	1
Mollusc	-	-	*Filopaludina Filopaludina doliaris*	2
Mollusc	-	-	*Filopaludina Filopaludina filosa*	3
Mollusc	-	-	*Filopaludina* sp.	2
Mollusc	-	-	*Mekongia* sp.	2
Mollusc	-	-	*Mekongia swainsoni braueri*	2
Mollusc	-	-	*Mekongia swainsoni flavida*	6
Mollusc	-	-	*Mekongia swainsoni swainsoni*	5
Mollusc	-	-	*Sinotaia aeruginosa*	3

Taxonomic group	**Order/Clade**	**Family**	**Species**	**Species occurrence**
Mollusc	-	-	*Trochotaia trochoides*	3
Mollusc	Heterobranchia (2)	Ellobiidae (2)	*Melampus nucleolus*	2
Mollusc	-	-	*Melampus fasciatulus*	2
Mollusc	Hygrophila (3)	Lymnaeidae (2)	*Lymnaea swinhoei*	1
Mollusc	-	-	*Lymnaea viridis*	11
Mollusc	-	Planorbidae	*Gyraulus* sp.	2
Mollusc	Heterostropha	Pyramidellidae	*Morrisonietta spiralis*	1
Mollusc	Neritimorpha (2)	Neritidae (2)	*Neritina rubida*	6
Mollusc	-	-	*Neritina violacea*	2
Mollusc	Unknown (2)	Unknown (2)	Gastropoda larva	6
Mollusc	-	-	Gastropoda sp.	1

Appendix T3

List of sampling sites with their LCBD indices, p-values and corrected p-values (cor-p). Corrected p-values that are significant at the 0.05 level are in bold (applied to the global communities). * indicates the hotspot location shared by different component communities. The first letter of each site name represents the country (T: Thailand, L: Laos, C: Cambodia, V: Vietnam); the second letter represents the channel of the river (M: Mekong, B: Bassac, T: Tributary). For example, TM1: site number 1 located on the Mekong in Thailand. R: richness (number of taxa).

Site		Global communities			Annelid communities			Crustacean communities			Mollusk communities			Insect communities		
code	R	LCBD	p	cor-p	LCBD	p	cor-p	LCBD	p	cor-p	LCBD	p	cor-p	LCBD	p	cor-p
CB1	30	0.0165	0.337	1	0.0153	0.516	1	0.0379	1E-04	0.006	0.0152	0.646	1	0.0186	0.201	1
CM1	6	0.0185	0.041	1	0.0165	0.414	1	0.0012	1	1	0.0142	0.771	1	0.0237	0.008	0.474
CM2	8	0.0174	0.166	1	0.0167	0.400	1	0.0012	1	1	0.0224	1E-04	0.006	0.0098	0.991	1
CM3*	12	0.0228	1E-04	**0.006**	0.0256	5E-04	0.032	0.0376	1E-04	0.006	0.0228	1E-04	0.006	0.0265	1E-04	0.006
CM4	16	0.0141	0.874	1	0.0120	0.833	1	0.0354	1E-04	0.006	0.0144	0.752	1	0.0158	0.477	1
CT1	11	0.0114	1	1	0.0165	0.422	1	0.0012	1	1	0.0115	0.986	1	0.0108	0.967	1
CT2	15	0.0124	0.993	1	0.0165	0.411	1	0.0012	1	1	0.0134	0.850	1	0.0120	0.891	1
CT3	26	0.0132	0.968	1	0.0117	0.853	1	0.0368	1E-04	0.006	0.0123	0.933	1	0.0136	0.733	1
CT4	10	0.0149	0.737	1	0.0167	0.411	1	0.0012	1	1	0.0131	0.875	1	0.0167	0.378	1
CT5	16	0.0142	0.852	1	0.0116	0.858	1	0.0012	1	1	0.0169	0.358	1	0.0154	0.535	1
CT6	6	0.0130	0.978	1	0.0165	0.414	1	0.0012	1	1	0.0231	1E-04	0.006	0.0103	0.982	1
CT7	25	0.0172	0.202	1	0.0131	0.766	1	0.0385	1E-04	0.006	0.0228	1E-04	0.006	0.0146	0.623	1
CT8	16	0.0191	0.014	0.756	0.0125	0.814	1	0.0394	1E-04	0.006	0.0164	0.427	1	0.0231	0.012	0.696
CT9	11	0.0122	0.997	1	0.0165	0.422	1	0.0012	1	1	0.0110	0.997	1	0.0121	0.878	1
CT10	17	0.0100	1	1	0.0119	0.849	1	0.0012	1	1	0.0107	0.999	1	0.0088	1	1
CT11	25	0.0112	1	1	0.0086	0.960	1	0.0394	1E-04	0.006	0.0118	0.969	1	0.0112	0.939	1
CT12	29	0.0139	0.894	1	0.0135	0.732	1	0.0386	1E-04	0.006	0.0097	1	1	0.0123	0.866	1
LM1	39	0.0159	0.484	1	0.0118	0.851	1	0.0428	1E-04	0.006	0.0189	0.065	1	0.0146	0.62	1
LM2	16	0.0150	0.707	1	0.0147	0.585	1	0.0012	1	1	0.0155	0.602	1	0.0158	0.48	1
LM3	19	0.0148	0.743	1	0.0140	0.674	1	0.0012	1	1	0.0135	0.843	1	0.0154	0.528	1
LM4	7	0.0146	0.790	1	0.0052	0.965	1	0.0012	1	1	0.0107	1	1	0.0139	0.713	1
LM5	24	0.0129	0.980	1	0.0116	0.859	1	0.0012	1	1	0.0201	0.007	0.320	0.0083	1	1
LM6	23	0.0186	0.034	1	0.0192	0.224	1	0.0424	1E-04	0.006	0.0209	4E-04	0.022	0.0190	0.161	1
LM7	8	0.0131	0.972	1	0.0167	0.405	1	0.0012	1	1	0.0042	1	1	0.0115	0.932	1
LT1	41	0.0154	0.615	1	0.0147	0.582	1	0.0012	1	1	0.0197	0.022	1.000	0.0124	0.868	1
LT2	24	0.0197	0.004	0.205	0.0208	0.159	1	0.0012	1	1	0.0224	1E-04	0.006	0.0174	0.309	1
LT3	15	0.0150	0.718	1	0.0244	0.009	0.527	0.0428	1E-04	0.006	0.0129	0.891	1	0.0152	0.55	1
LT4	25	0.0118	0.998	1	0.0128	0.786	1	0.0012	1	1	0.0119	0.965	1	0.0116	0.918	1
LT5	6	0.0113	1	1	0.0052	0.964	1	0.0012	1	1	0.0107	0.999	1	0.0101	0.986	1
LT6	31	0.0103	1	1	0.0113	0.875	1	0.0012	1	1	0.0105	1	1	0.0087	0.999	1

Site		**Global communities**			**Annelid communities**			**Crustacean communities**			**Mollusk communities**			**Insect communities**		
code	**R**	**LCBD**	**p**	**cor-p**	**LCBD**	**p**	**cor-p**	**LCBD**	**p**	**cor-p**	**LCBD**	**p**	**cor-p**	**LCBD**	**p**	**cor-p**
LT7	28	0.0121	0.997	1	0.0126	0.800	1	0.0422	1E-04	0.006	0.0185	0.104	1	0.0100	0.99	1
LT8	31	0.0152	0.669	1	0.0162	0.435	1	0.0394	1E-04	0.006	0.0136	0.825	1	0.0138	0.716	1
LT9	12	0.0137	0.911	1	0.0052	0.966	1	0.0012	1	1	0.0042	1	1	0.0112	0.949	1
LT10	19	0.0175	0.156	1	0.0244	0.009	0.561	0.0012	1	1	0.0185	0.098	1	0.0176	0.29	1
LT11	41	0.0138	0.915	1	0.0198	0.201	1	0.0012	1	1	0.0154	0.618	1	0.0130	0.806	1
LT12	32	0.0152	0.656	1	0.0169	0.387	1	0.0012	1	1	0.0175	0.247	1	0.0152	0.558	1
TM1	9	0.0200	1E-04	**0.012**	0.0192	0.236	1	0.0012	1	1	0.0042	1	1	0.0233	0.012	0.702
TM2	21	0.0210	1E-04	**0.006**	0.0192	0.239	1	0.0379	1E-04	0.006	0.0216	1E-04	0.006	0.0238	0.005	0.317
TM3	12	0.0158	0.519	1	0.0165	0.414	1	0.0012	1	1	0.0107	0.999	1	0.0165	0.406	1
TM4	12	0.0209	1E-04	**0.006**	0.0192	0.236	1	0.0012	1	1	0.0207	0.001	0.065	0.0207	0.073	1
TM5	12	0.0146	0.785	1	0.0165	0.416	1	0.0012	1	1	0.0130	0.877	1	0.0153	0.542	1
TT1	18	0.0165	0.353	1	0.0158	0.472	1	0.0012	1	1	0.0198	0.016	0.750	0.0179	0.262	1
TT2	10	0.0153	0.630	1	0.0165	0.412	1	0.0012	1	1	0.0211	2E-04	0.011	0.0146	0.625	1
TT3	16	0.0176	0.136	1	0.0221	0.107	1	0.0012	1	1	0.0102	0.999	1	0.0206	0.075	1
TT4	8	0.0133	0.953	1	0.0139	0.682	1	0.0012	1	1	0.0218	1E-04	0.006	0.0109	0.962	1
TT5	10	0.0200	0.001	0.065	0.0192	0.242	1	0.0012	1	1	0.0208	9E-04	0.047	0.0214	0.051	1
TT6	23	0.0165	0.341	1	0.0192	0.233	1	0.0379	1E-04	0.006	0.0202	0.007	0.312	0.0162	0.444	1
TT7	27	0.0144	0.823	1	0.0130	0.775	1	0.0012	1	1	0.0137	0.831	1	0.0105	0.973	1
TT8	17	0.0188	0.027	1	0.0192	0.239	1	0.0012	1	1	0.0208	0.001	0.051	0.0199	0.119	1
VB1	74	0.0162	0.420	1	0.0190	0.252	1	0.0341	1E-04	0.006	0.0135	0.834	1	0.0189	0.182	1
VB2	18	0.0168	0.286	1	0.0128	0.791	1	0.0353	1E-04	0.006	0.0167	0.384	1	0.0181	0.239	1
VB3	52	0.0195	0.007	0.226	0.0196	0.210	1	0.0354	1E-04	0.006	0.0197	0.025	1.000	0.0204	0.095	1
VB4	34	0.0204	5E-04	**0.012**	0.0218	0.114	1	0.0388	1E-04	0.006	0.0189	0.069	1	0.0248	0.002	0.124
VB5	66	0.0157	0.533	1	0.0189	0.247	1	0.0356	1E-04	0.006	0.0141	0.78	1	0.0157	0.495	1
VM1	53	0.0189	0.017	0.896	0.0142	0.639	1	0.0360	1E-04	0.006	0.0195	0.032	1	0.0197	0.129	1
VM2	67	0.0177	0.128	1	0.0186	0.269	1	0.0360	1E-04	0.006	0.0160	0.519	1	0.0175	0.301	1
VM3	19	0.0166	0.318	1	0.0135	0.728	1	0.0400	1E-04	0.006	0.0125	0.92	1	0.0144	0.648	1
VM4	33	0.0193	0.009	0.468	0.0212	0.140	1	0.0390	1E-04	0.006	0.0186	0.087	1	0.0218	0.038	1
VM5	61	0.0176	0.126	1	0.0180	0.310	1	0.0358	1E-04	0.006	0.0180	0.165	1	0.0185	0.212	1
VM6	49	0.0190	0.002	0.104	0.0192	0.242	1	0.0012	1	1	0.0202	0.006	0.27	0.0218	0.04	1
VT1	19	0.0190	0.017	0.896	0.0117	0.856	1	0.0012	1	1	0.0208	5E-04	0.027	0.0148	0.599	1
VT2	10	0.0140	0.882	1	0.0158	0.467	1	0.0012	1	1	0.0107	0.999	1	0.0150	0.568	1
VT3	8	0.0165	0.344	1	0.0165	0.415	1	0.0012	1	1	0.0214	1E-04	0.006	0.0168	0.368	1

Appendix T4

List of taxa and their occurrences recorded in Flemish rivers from 1990 to 2010. The numbers of the occurrence and abundance of each taxon were calculated based on the median values of each period. D1: 1991-1995, D2: 1996-2000, D3: 2001-2005, and D4: 2006-2010.

Taxonomic				**Taxon**	**Occurrence**				**Abundance**			
groups	**Oder/clade**	**Family**	**Taxon**	**Code**	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Annelida.1	Arhynchobdellida.1	Erpobdellidae.1	*Dina*	A01	0	1	1	0	0	1	4	0
Annelida.2	Arhynchobdellida.2	Erpobdellidae.2	*Erpobdella*	A02	621	1,077	1,530	1,026	3,939	7,431	13,625	25,202
Annelida.3	Arhynchobdellida.3	Erpobdellidae.3	*Trocheta*	A03	2	11	24	26	1	14	57	73
Annelida.4	Arhynchobdellida.4	Haemopidae	*Haemopis*	A04	31	44	51	26	45	332	151	54
Annelida.5	Arhynchobdellida.5	Hirudinidae	*Hirudo*	A05	2	3	1	2	3	4	1	11
Annelida.6	Branchiobdellida	Branchiobdellidae	Branchiobdellidae	A06	4	2	0	1	5	2	0	3
Annelida.7	Glossiphoniidae	Glossiphoniidae	*Theromyzon*	A07	141	287	508	386	214	487	887	974
Annelida.8	Haplotaxida.1	Enchytraeidae	Enchytraeidae	A08	231	84	6	0	16,589	947	107	0
Annelida.9	Haplotaxida.2	Haplotaxidae	Haplotaxidae	A09	5	1	0	0	25,551	1	0	0
Annelida.10	Haplotaxida.3	Lumbricidae	Lumbricidae	A10	161	312	401	257	336	605	915	668
Annelida.11	Haplotaxida.4	Naididae	Naididae	A11	1292	2,020	2,595	1,699	243,998	238,425	225,592	329,684
Annelida.12	Lumbriculida	Lumbriculidae	Lumbriculidae	A12	193	274	363	186	1,263	875	3,062	1,996
Annelida.13	Rhynchobdellida.1	Glossiphoniidae.1	*Glossiphonia*	A13	555	1,041	1,572	957	2,259	4,621	13,646	22,299
Annelida.14	Rhynchobdellida.2	Glossiphoniidae.2	*Haementeria*	A14	7	9	3	0	10	12	3	0
Annelida.15	Rhynchobdellida.3	Glossiphoniidae.3	*Helobdella*	A15	787	1,299	1,782	1,182	7,082	11,161	26,205	43,666
Annelida.16	Rhynchobdellida.4	Glossiphoniidae.4	*Hemiclepsis*	A16	85	199	352	223	165	319	930	560
Annelida.17	Rhynchobdellida.5	Piscicolidae.1	*Cystobranchus*	A17	0	0	1	0	0	0	1	0
Annelida.18	Rhynchobdellida.6	Piscicolidae.2	*Piscicola*	A18	148	255	386	215	338	812	1,504	1,029
Annelida.19	Aelosomatida	Aelosomatidae	Aelosomatidae	A19	1	0	0	0	1	0	0	0
Annelida.20	Canalipalpata	Ampharetidae	Ampharetidae	A20	0	0	11	42	0	0	912	1,105
Annelida.21	NA	NA	Polychaeta	A21	0	0	20	0	0	0	145	0
Crustacea.1	Cladocera	Daphniidae	*Daphnia*	C01	1	0	0	0	1	0	0	0
Crustacea.2	Anostraca	Chirocephalidae	Chirocephalidae	C02	0	0	1	0	0	0	11	0
Crustacea.3	Spinicaudata.1	Leptestheriidae	Leptestheriidae	C03	0	1	1	152	0	6	1	1,562
Crustacea.4	Spinicaudata.2	Limnadiidae	Limnadiidae	C04	2	2	17	137	3	3	52	3,963
Crustacea.5	Amphipoda.1	Corophiidae	Corophiidae	C05	40	53	62	84	3,652	1,803	1,357	16,873
Crustacea.6	Amphipoda.2	Crangonyctidae	Crangonyctidae	C06	2	7	83	62	13	11	992	2,151
Crustacea.7	Amphipoda.3	Gammaridae	Gammaridae	C07	525	850	1,300	816	15,314	24,577	84,008	199,076
Crustacea.8	Amphipoda.4	Talitridae	Talitridae	C08	19	19	16	13	84	85	56	26
Crustacea.9	Decapoda.1	Palaemonidae	Palaemonidae	C09	106	47	93	35	1,658	373	1,826	637
Crustacea.10	Decapoda.2	Panopeidae	Panopeidae	C10	0	0	0	2	0	0	0	7
Crustacea.11	Decapoda.3	Astacidae	Astacidae	C11	8	4	1	6	64	34	11	351
Crustacea.12	Decapoda.4	Atyidae	Atyidae	C12	19	33	18	10	62	122	28	44

Taxonomic				**Taxon**	**Occurrence**				**Abundance**			
groups	**Oder/clade**	**Family**	**Taxon**	**Code**	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Crustacea.13	Decapoda.5	Cambaridae	Cambaridae	C13	16	32	69	59	23	136	163	164
Crustacea.14	Decapoda.6	Grapsidae	Grapsidae	C14	2	2	5	9	3	53	70	32
Crustacea.15	Isopoda.1	Asellidae	Asellidae	C15	1164	1,762	2,345	1,366	46,208	60,840	200,461	210,784
Crustacea.16	Isopoda.2	Janiridae	Janiridae	C16	0	0	9	20	0	0	249	2,507
Crustacea.17	Isopoda.3	Sphaeromatidae	Sphaeromatidae	C17	0	0	1	3	0	0	1	84
Crustacea.18	Mysida	Mysidae	Mysidae	C18	91	64	77	61	7,854	6,109	32,815	102,156
Crustacea.19	Arguloida	Argulidae	Argulidae	C19	5	25	48	24	55	84	647	324
Crustacea.20	NA	NA	Copepoda	C20	0	1	0	0	0	1	0	0
Crustacea.21	NA	NA	Ostracoda	C21	226	219	337	116	612	946	44,647	15,549
Insecta.1	Coleoptera.1	Dryopidae	Dryopidae	I001	15	18	77	318	20	18	143	4,187
Insecta.2	Coleoptera.2	Dytiscidae	Dytiscidae	I002	653	882	982	286	2,118	2,919	4,513	1,318
Insecta.3	Coleoptera.3	Elminthidae	Elminthidae	I003	32	60	103	49	67	113	624	281
Insecta.4	Coleoptera.4	Gyrinidae	Gyrinidae	I004	11	22	47	47	14	45	118	324
Insecta.5	Coleoptera.5	Haliplidae	Haliplidae	I005	261	400	549	387	612	1,114	2,840	2,496
Insecta.6	Coleoptera.6	Hydraenidae	Hydraenidae	I006	34	17	53	18	48	27	84	28
Insecta.7	Coleoptera.7	Hydrophilidae	Hydrophilidae	I007	197	333	709	333	349	781	3,049	1,176
Insecta.8	Coleoptera.8	Hygrobiidae	Hygrobiidae	I008	15	18	26	8	34	73	39	25
Insecta.9	Coleoptera.9	Noteridae	Noteridae	I009	49	94	150	57	160	222	468	159
Insecta.10	Coleoptera.10	Psephenidae	Psephenidae	I010	2	1	1	1	12	51	1	2
Insecta.11	Coleoptera.11	Scirtidae	Scirtidae	I011	18	67	157	37	72	411	1,215	167
Insecta.12	Diptera.1	Athericidae	Athericidae	I012	3	9	25	4	5	23	51	9
Insecta.13	Diptera.2	Blephariceridae	Blephariceridae	I013	2	2	0	0	7	4	0	0
Insecta.14	Diptera.3	Ceratopogonidae	Ceratopogonidae	I014	120	158	515	297	214	259	2,972	1,675
Insecta.15	Diptera.4	Chaoboridae	Chaoboridae	I015	11	19	26	31	168	159	75	531
Insecta.16	Diptera.5	Chironomidae.1	Chironomidae. non *thummi-plumosus*	I016	1289	1,922	2,586	1,651	37,945	60,515	133,648	201,377
Insecta.17	Diptera.6	Chironomidae.2	Chironomidae. *thummi-plumosus*	I017	983	1,371	1,670	1,154	36,334	38,999	98,689	186,934
Insecta.18	Diptera.7	Culicidae	Culicidae	I018	110	228	357	264	1,714	974	2,223	3,831
Insecta.19	Diptera.8	Cylindrotomidae	Cylindrotomidae	I019	6	1	11	3	7	1	26	14
Insecta.20	Diptera.9	Dixidae	Dixidae	I020	5	33	44	22	8	71	125	25
Insecta.21	Diptera.10	Dolichopodidae	Dolichopodidae	I021	5	17	29	7	4	17	35	7
Insecta.22	Diptera.11	Empididae	Empididae	I022	18	25	81	46	51	31	155	97
Insecta.23	Diptera.12	Ephydridae	Ephydridae	I023	5	21	111	34	5	48	176	67
Insecta.24	Diptera.13	Muscidae	Muscidae	I024	17	32	102	69	29	59	159	98
Insecta.25	Diptera.14	Psychodidae	Psychodidae	I025	165	292	731	213	324	678	2,821	1,801
Insecta.26	Diptera.15	Ptychopteridae	Ptychopteridae	I026	29	54	75	9	70	389	327	72

Taxonomic				**Taxon**	**Occurrence**				**Abundance**			
groups	**Oder/clade**	**Family**	**Taxon**	**Code**	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Insecta.27	Diptera.16	Rhagionidae	Rhagionidae	I027	4	15	40	29	5	14	44	31
Insecta.28	Diptera.17	Scatophagidae	Scatophagidae	I028	2	2	7	4	2	52	16	4
Insecta.29	Diptera.18	Sciomyzidae	Sciomyzidae	I029	21	25	86	53	82	36	159	123
Insecta.30	Diptera.19	Simuliidae	Simuliidae	I030	182	494	794	555	2,664	6,398	31,695	31,620
Insecta.31	Diptera.20	Stratiomyidae	Stratiomyidae	I031	51	126	311	131	66	144	543	285
Insecta.32	Diptera.21	Syrphidae	Eristalinae	I032	22	38	74	64	27	48	158	293
Insecta.33	Diptera.22	Tabanidae	Tabanidae	I033	69	68	114	40	143	112	231	111
Insecta.34	Diptera.23	Thaumaleidae	Thaumaleidae	I034	1	2	2	1	11	2	3	1
Insecta.35	Diptera.24	Tipulidae.1	Limoniidae	I035	175	312	480	58	432	1,131	1,842	290
Insecta.36	Diptera.25	Tipulidae.2	Tipulidae	I036	181	272	449	275	301	432	932	724
Insecta.37	Ephemeroptera.1	Baetidae.1	*Baetis*	I037	225	535	722	451	2,293	7,776	12,359	16,359
Insecta.38	Ephemeroptera.2	Baetidae.2	*Centroptilum*	I038	6	5	2	1	6	17	9	65
Insecta.39	Ephemeroptera.3	Baetidae.3	*Cloeon*	I039	373	547	743	416	4,746	5,852	10,095	21,110
Insecta.40	Ephemeroptera.4	Baetidae.4	*Procloeon*	I040	0	1	2	1	0	1	5	4
Insecta.41	Ephemeroptera.5	Caenidae.1	*Brachycercus*	I041	1	3	2	0	2	4	12	0
Insecta.42	Ephemeroptera.6	Caenidae.2	*Caenis*	I042	102	158	288	202	719	873	2,661	3,716
Insecta.43	Ephemeroptera.7	Ephemerellidae	*Ephemerella*	I043	1	7	12	3	1	25	31	45
Insecta.44	Ephemeroptera.8	Ephemeridae	*Ephemera*	I044	18	24	38	21	71	113	137	146
Insecta.45	Ephemeroptera.9	Heptageniidae.1	*Ecdyonurus*	I045	1	1	1	0	2	1	3	0
Insecta.46	Ephemeroptera.10	Heptageniidae.2	*Epeorus*	I046	0	0	0	1	0	0	0	1
Insecta.47	Ephemeroptera.11	Heptageniidae.3	*Heptagenia*	I047	12	41	37	13	137	301	276	121
Insecta.48	Ephemeroptera.12	Leptophlebiidae.1	*Habrophlebia*	I048	0	3	1	0	0	4	6	0
Insecta.49	Ephemeroptera.13	Leptophlebiidae.2	*Leptophlebia*	I049	0	0	1	0	0	0	1	0
Insecta.50	Ephemeroptera.14	Leptophlebiidae.3	*Paraleptophlebia*	I050	0	1	2	0	0	3	32	0
Insecta.51	Ephemeroptera.15	Polymitarcyidae	*Ephoron*	I051	0	0	0	1	0	0	0	1
Insecta.52	Ephemeroptera.16	Potamanthidae.1	*Potamanthus*	I052	0	0	1	0	0	0	9	0
Insecta.53	Ephemeroptera.17	Potamanthidae.2	*Potamopyrgus*	I053	204	280	463	348	3,911	4,256	30,853	42,503
Insecta.54	Hemiptera.1	Aphelocheiridae	*Aphelocheirus*	I054	17	23	25	21	48	81	121	119
Insecta.55	Hemiptera.2	Corixidae.1	*Arctocorisa*	I055	5	12	0	0	8	27	0	0
Insecta.56	Hemiptera.3	Corixidae.2	*Callicorixa*	I056	27	33	28	7	45	35	61	10
Insecta.57	Hemiptera.4	Corixidae.3	*Corixa*	I057	159	189	168	65	497	499	487	232
Insecta.58	Hemiptera.5	Corixidae.4	*Cymatia*	I058	4	18	31	13	7	215	262	195
Insecta.59	Hemiptera.6	Corixidae.5	*Glaenocorisa*	I059	0	1	0	0	0	1	0	0
Insecta.60	Hemiptera.7	Corixidae.6	*Hesperocorixa*	I060	83	155	164	90	257	264	426	346
Insecta.61	Hemiptera.8	Corixidae.7	*Micronecta*	I061	38	85	220	141	385	2,085	3,506	10,195
Insecta.62	Hemiptera.9	Corixidae.8	*Paracorixa*	I062	19	8	8	2	24	31	37	2
Insecta.63	Hemiptera.10	Corixidae.9	*Sigara*	I063	633	826	978	595	10,525	10,253	18,401	13,726

Taxonomic groups	**Oder/clade**	**Family**	**Taxon**	**Taxon Code**	**Occurrence**				**Abundance**			
					D1	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Insecta.64	Hemiptera.11	Gerridae	*Gerris*	I064	61	144	298	173	133	318	2,369	2,161
Insecta.65	Hemiptera.12	Hebridae	*Hebrus*	I065	1	1	0	0	1	1	0	0
Insecta.66	Hemiptera.13	Hydrometridae	*Hydrometra*	I066	8	10	34	11	11	22	61	21
Insecta.67	Hemiptera.14	Mesoveliidae	*Mesovelia*	I067	7	4	6	2	6	6	8	6
Insecta.68	Hemiptera.15	Naucoridae	*llyocoris*	I068	53	61	130	67	133	201	566	424
Insecta.69	Hemiptera.16	Naucoridae	*Naucoris*	I069	4	33	13	6	5	119	98	53
Insecta.70	Hemiptera.17	Nepidae.1	*Nepa*	I070	45	109	156	73	58	144	242	127
Insecta.71	Hemiptera.18	Nepidae.2	*Ranatra*	I071	26	27	48	19	202	151	96	85
Insecta.72	Hemiptera.19	Notonectidae	*Notonecta*	I072	157	266	393	237	232	455	1,119	604
Insecta.73	Hemiptera.20	Pleidae	*Plea*	I073	30	42	104	79	65	149	581	4,114
Insecta.74	Hemiptera.21	Veliidae.1	*Microvelia*	I074	4	3	19	4	3	8	39	36
Insecta.75	Hemiptera.22	Veliidae.2	*Velia*	I075	22	74	147	97	69	225	489	296
Insecta.76	Megaloptera	Sialidae	*Sialis*	I076	182	276	277	173	672	835	985	871
Insecta.77	Neuroptera	NA	*Neuroptera*	I077	2	6	4	0	2	58	8	0
Insecta.78	Odonata.1	Aeshnidae.1	*Anax*	I078	19	14	39	21	19	19	74	38
Insecta.79	Odonata.2	Aeshnidae.2	*Brachytron*	I079	2	1	1	1	2	1	1	1
Insecta.80	Odonata.3	Aeshnidae.3	*Aeschna*	I080	23	42	96	34	28	43	138	58
Insecta.81	Odonata.4	Calopterygidae	*Calopteryx*	I081	54	84	189	157	106	171	934	1,449
Insecta.82	Odonata.5	Coenagrionidae.1	*Cercion*	I082	5	2	1	4	6	2	1	57
Insecta.83	Odonata.6	Coenagrionidae.2	*Ceriagrion*	I083	2	1	3	1	3	1	5	1
Insecta.84	Odonata.7	Coenagrionidae.3	*Coenagrion*	I084	30	93	191	92	69	396	840	629
Insecta.85	Odonata.8	Coenagrionidae.4	*Enallagma*	I085	0	3	4	0	0	15	14	0
Insecta.86	Odonata.9	Coenagrionidae.5	*Erythromma*	I086	5	8	42	20	8	11	309	138
Insecta.87	Odonata.10	Coenagrionidae.6	*Ischnura*	I087	376	579	706	466	2,117	4,012	6,298	8,023
Insecta.88	Odonata.11	Coenagrionidae.7	*Nehalennia*	I088	2	4	0	1	7	5	0	1
Insecta.89	Odonata.12	Coenagrionidae.8	*Pyrrhosoma*	I089	7	10	47	21	12	14	301	82
Insecta.90	Odonata.13	Cordulegastridae	*Cordulegaster*	I090	5	5	10	3	5	6	11	6
Insecta.91	Odonata.14	Corduliidae.1	*Cordulia*	I091	1	0	1	1	1	0	1	1
Insecta.92	Odonata.15	Corduliidae.2	*Oxygastra*	I092	0	1	0	0	0	1	0	0
Insecta.93	Odonata.16	Corduliidae.3	*Somatochlora*	I093	0	0	0	1	0	0	0	7
Insecta.94	Odonata.17	Gomphidae.1	*Gomphus*	I094	7	13	9	3	13	34	73	7
Insecta.95	Odonata.18	Gomphidae.2	*Onychogomphus*	I095	0	0	1	0	0	0	1	0
Insecta.96	Odonata.19	Lestidae.1	*Lestes*	I096	18	66	69	37	39	168	546	347
Insecta.97	Odonata.20	Lestidae.2	*Sympecma*	I097	6	2	6	0	7	3	11	0
Insecta.98	Odonata.21	Libellulidae.1	*Crocothemis*	I098	0	0	2	1	0	0	4	1
Insecta.99	Odonata.22	Libellulidae.2	*Libellula*	I099	7	7	6	18	8	7	12	33
Insecta.100	Odonata.23	Libellulidae.3	*Orthetrum*	I100	7	21	18	26	9	24	23	59

Taxonomic				**Taxon**	**Occurrence**				**Abundance**			
groups	**Oder/clade**	**Family**	**Taxon**	**Code**	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Insecta.101	Odonata.24	Libellulidae.4	*Sympetrum*	I101	2	4	20	12	3	3	42	19
Insecta.102	Odonata.25	Platycnemididae	*Platycnemis*	I102	66	103	77	48	178	234	216	177
Insecta.103	Plecoptera.1	Capniidae	*Capnia*	I103	0	1	0	0	0	1	0	0
Insecta.104	Plecoptera.2	Leuctridae	*Leuctra*	I104	1	1	8	5	2	1	24	58
Insecta.105	Plecoptera.3	Nemouridae.1	*Amphinemoura*	I105	0	0	4	2	0	0	11	8
Insecta.106	Plecoptera.4	Nemouridae.2	*Nemoura*	I106	32	45	109	50	238	465	1,165	4,672
Insecta.107	Plecoptera.5	Nemouridae.3	*Nemourella*	I107	1	8	10	2	1	33	86	60
Insecta.108	Plecoptera.6	Nemouridae.4	*Protonemoura*	I108	0	5	7	1	0	87	74	12
Insecta.109	Plecoptera.7	Perlidae.1	*Marthamea*	I109	0	1	0	0	0	2	0	0
Insecta.110	Plecoptera.8	Perlidae.2	*Perla*	I110	0	0	0	1	0	0	0	1
Insecta.111	Plecoptera.9	Perlodidae.1	*Isogenus*	I111	0	0	0	1	0	0	0	1
Insecta.112	Plecoptera.10	Perlodidae.2	*Perlodes*	I112	0	1	1	0	0	1	4	0
Insecta.113	Plecoptera.11	Taeniopterygidae	*Rhabdiopteryx*	I113	0	1	0	0	0	1	0	0
Insecta.114	Trichoptera.1	Beraeidae	Beraeidae	I114	6	10	9	5	17	34	9	6
Insecta.115	Trichoptera.2	Brachycentridae	Brachycentridae	I115	3	0	2	0	3	0	1	0
Insecta.116	Trichoptera.3	Ecnomidae	Ecnomidae	I116	78	93	106	30	399	496	503	203
Insecta.117	Trichoptera.4	Glossosomatidae	Glossosomatidae	I117	2	2	3	3	2	2	22	10
Insecta.118	Trichoptera.5	Goeridae	Goeridae	I118	5	16	19	8	19	55	97	55
Insecta.119	Trichoptera.6	Hydropsychidae	Hydropsychidae	I119	103	164	342	264	528	1,241	3,708	7,896
Insecta.120	Trichoptera.7	Hydroptilidae	Hydroptilidae	I120	43	132	242	170	164	664	2,261	3,716
Insecta.121	Trichoptera.8	Lepidostomatidae	Lepidostomatidae	I121	9	4	10	0	10	6	33	0
Insecta.122	Trichoptera.9	Leptoceridae	Leptoceridae	I122	149	211	343	104	595	927	2,650	851
Insecta.123	Trichoptera.10	Limnephilidae	Limnephilidae	I123	104	179	316	190	237	717	1,944	770
Insecta.124	Trichoptera.11	Molannidae	Molannidae	I124	6	6	11	21	9	7	21	49
Insecta.125	Trichoptera.12	Odontoceridae	Odontoceridae	I125	1	1	1	0	1	1	2	0
Insecta.126	Trichoptera.13	Philopotamidae	Philopotamidae	I126	2	2	7	0	3	2	13	0
Insecta.127	Trichoptera.14	Phryganeidae	Phryganeidae	I127	17	20	36	34	41	36	69	56
Insecta.128	Trichoptera.15	Polycentropodidae	Polycentropodidae	I128	86	154	248	110	406	707	1,075	857
Insecta.129	Trichoptera.16	Psychomyidae	Psychomyidae	I129	17	39	100	71	22	73	244	272
Insecta.130	Trichoptera.17	Rhyacophilidae	Rhyacophilidae	I130	0	16	16	6	0	67	147	61
Insecta.131	Trichoptera.18	Sericostomatidae	Sericostomatidae	I131	10	17	70	20	12	23	313	108
Mollusca.1	Unionida.1	Margaritiferidae	*Margaritifera*	M01	0	1	0	0	0	1	0	0
Mollusca.2	Unionida.2	Unionidae.1	*Pseudanodonta*	M02	2	1	1	55	2	1	1	168
Mollusca.3	Unionida.3	Unionidae.2	*Unio*	M03	5	3	8	13	9	4	27	75
Mollusca.4	Unionida.4	Unionidae.3	*Anodonta*	M04	8	12	32	19	11	16	130	82
Mollusca.5	**Veneroida.1**	**Corbiculidae**	***Corbicula***	**M05**	**0**	**12**	**34**	**62**	**0**	**25**	**144**	**983**
Mollusca.6	**Veneroida.2**	**Dreissenidae.1**	***Dreissena***	**M06**	**61**	**101**	**168**	**196**	**1,124**	**1,151**	**2,309**	**26,421**

Taxonomic groups	Oder/clade	Family	Taxon	Taxon Code	Occurrence				Abundance			
					D1	D2	D3	D4	D1	D2	D3	D4
Mollusca.7	Veneroida.3	Dreissenidae.2	*Mytilopsis*	M07	0	1	0	0	0	2	0	0
Mollusca.8	Veneroida.4	Sphaeriidae.1	*Pisidium*	M08	469	732	1,135	765	7,901	11,699	32,722	43,344
Mollusca.9	Veneroida.5	Sphaeriidae.2	*Sphaerium*	M09	253	378	696	464	2,682	2,529	5,364	7,836
Mollusca.10	Caenogastropoda.1	Amnicolidae.1	*Bythinella*	M10	3	3	0	0	18	15	0	0
Mollusca.11	Caenogastropoda.2	Amnicolidae.2	*Marstoniopsis*	M11	1	1	0	0	1	1	0	0
Mollusca.12	Caenogastropoda.3	Bithyniidae	*Bithynia*	M12	277	444	683	453	2,867	5,588	23,472	21,672
Mollusca.13	Caenogastropoda.4	Hydrobiidae	*Pseudamnicola*	M13	1	7	2	352	11	46	3	3,665
Mollusca.14	**Caenogastropoda.5**	**Lithoglyphidae**	***Lithoglyphus***	**M14**	**7**	**5**	**5**	**3**	**40**	**18**	**14**	**6**
Mollusca.15	**Caenogastropoda.6**	**Viviparidae**	***Viviparus***	**M15**	**27**	**22**	**27**	**15**	**167**	**76**	**103**	**132**
Mollusca.16	Hygrophila.1	Acroloxidae	*Acroloxus*	M16	45	88	214	197	126	227	1,261	1,035
Mollusca.17	Hygrophila.2	Lymnaeidae.1	*Lymnaea*	M17	704	1,133	1,507	862	17,066	9,022	22,229	15,933
Mollusca.18	Hygrophila.3	Lymnaeidae.2	*Myxas*	M18	1	1	0	0	2	1	0	0
Mollusca.19	Hygrophila.4	Physidae.1	*Aplexa*	M19	14	8	14	2	49	29	126	3
Mollusca.20	Hygrophila.5	Physidae.2	*Physa*	M20	640	946	1,072	207	9,158	11,818	11,949	3,725
Mollusca.21	**Hygrophila.6**	**Physidae.3**	***Physella***	**M21**	**1**	**1**	**227**	**754**	**7**	**1**	**11,155**	**35,560**
Mollusca.22	Hygrophila.7	Planorbidae.1	*Ancylus*	M22	42	94	159	126	91	444	685	1,113
Mollusca.23	Hygrophila.8	Planorbidae.2	*Anisus*	M23	213	306	438	221	879	1,872	5,483	2,019
Mollusca.24	Hygrophila.9	Planorbidae.3	*Armiger*	M24	38	78	269	139	116	333	1,669	756
Mollusca.25	Hygrophila.10	Planorbidae.4	*Bathyomphalus*	M25	104	185	239	155	486	867	1,637	1,544
Mollusca.26	**Hygrophila.11**	**Planorbidae.5**	***Ferrisia***	**M26**	**7**	**20**	**160**	**108**	**58**	**49**	**788**	**842**
Mollusca.27	Hygrophila.12	Planorbidae.6	*Gyraulus*	M27	274	429	571	450	1,609	2,539	4,858	8,757
Mollusca.28	Hygrophila.13	Planorbidae.7	*Hippeutis*	M28	22	64	201	130	92	179	1,243	945
Mollusca.29	**Hygrophila.14**	**Planorbidae.8**	***Menetus***	**M29**	**0**	**0**	**0**	**3**	**0**	**0**	**0**	**8**
Mollusca.30	Hygrophila.15	Planorbidae.9	*Planorbarius*	M30	110	153	208	150	945	622	1,090	967
Mollusca.31	Hygrophila.16	Planorbidae.10	*Planorbis*	M31	164	216	275	194	1,183	1,606	1,662	2,172
Mollusca.32	Hygrophila.17	Planorbidae.11	*Segmentina*	M32	27	42	100	60	64	148	832	449
Mollusca.33	Heterobranchia	Valvatidae	*Valvata*	M33	302	404	666	430	5,494	7,328	29,679	39,508
Mollusca.34	Neritimorpha	Neritidae	*Theodoxus*	M34	13	16	2	4	90	73	3	71

Note: Mollusc taxa in **bold** are alien genera which were used in the predictions in the case study of Flemish rivers.

PART II: PUBLICATIONS

Limnologica 64 (2017) 20–30

Contents lists available at ScienceDirect

Limnologica

journal homepage: www.elsevier.com/locate/limno

Spatial organization of macroinvertebrate assemblages in the Lower Mekong Basin

Ratha Sor[a,b,c,*], Pieter Boets[b,d], Ratha Chea[a], Peter L.M. Goethals[b], Sovan Lek[a]

[a] Laboratoire Evolution & Diversité Biologique, UMR 5174, Université Paul Sabatier – Toulouse III, 118 route de Narbonne, 31062 Toulouse cédex 4, France
[b] Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure links 653, B9000 Ghent, Belgium
[c] Department of Biology, Faculty of Science, Royal University of Phnom Penh,Russian Boulevard, 12000, Phnom Penh, Cambodia
[d] Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000 Ghent, Belgium

ARTICLE INFO

Keywords:
Spatial patterns
Indicator species
Land cover
Watershed surface area
River ecology
Biodiversity hotspot

ABSTRACT

In this study, we investigated patterns of spatial variation in macroinvertebrate assemblages in the Lower Mekong Basin (LMB) and examined their relationship with environmental factors. Cluster analysis was used to group macroinvertebrate samples and Linear Discriminant Analysis was performed to discriminate the major factors associated with the macroinvertebrate assemblages. Four clusters could be distinguished based on the dissimilarity between macroinvertebrate assemblages. The assemblages related to the tributaries and the upstream parts (cluster II) were characterized by a lower richness, abundance, diversity and a lower number of indicator taxa compared to the assemblage found downstream in the Mekong delta (cluster I). Aquatic insects and their indicator taxa (e.g. *Caenodes* sp., *Dipseudopsis* sp. and Gomphidae sp.), preferring a high-altitude environment with a high dissolved oxygen concentration and a high density of wood/shrub and evergreen forests, were the most predominant group in the assemblages occupying the tributaries and the upstream parts (cluster IIa). The assemblage found in the delta, consisting largely of molluscs and a moderate richness and abundance of worms, crustaceans and dipteran insects, was mainly represented by *Corbicula leviuscula* and *C. moreletiana* (molluscs), *Namalycastis longicirris* and *Chaetogaster langi* (worms), *Corophium minutum* and *Grandidierella lignorum* (crustaceans), and *Cricotopus* sp. and *Clinotanypus* sp. (dipteran insects). This assemblage was associated with a large watershed surface area, deep and wide rivers and a high water temperature. The intermediate assemblage (cluster IIb1) in-between could be discriminated based on land cover types including inundated, wetland and agricultural land, and was represented most by molluscs. Strikingly, the assemblage occupying the upstream parts (cluster IIa), which is related to intensified agriculture and a moderate conductivity, was characterized by a higher macroinvertebrate diversity compared to the mountainous and less impacted tributaries. This could mean that the natural stress is high in these systems for some taxa, leading to a lower overall taxonomic richness and abundance. Nevertheless, the number of taxa and the diversity of macroinvertebrates remained relatively high across the basin, especially in the delta assemblage. Therefore, the LMB deserves a particular attention for conservation.

1. Introduction

Tropical regions are remarkably rich in biodiversity (Sodhi et al., 2004), with 16 out of the 25 terrestrial biodiversity hotspots of the world being located in the tropical zone (Myers et al., 2000). Southeast Asia has a unique geological history (Sodhi et al., 2004), and can be separated into four biodiversity hotspot regions: The Philippines, Sundaland, Wallacea and Indo-Burma (Mittermeier et al., 1999). Through Indo-Burma, the Mekong River, which is the longest river in Southeast Asia, flows from its source in the Tibetan plateau to the South China Sea (Zalinge et al., 2003). This river harbours diverse biotic communities and is the breeding area of numerous endemic, threatened and endangered species. The aquatic and terrestrial communities of the Mekong form a biologically important food web that supports a high biodiversity (Sodhi et al., 2004). Moreover, it is an economically important region, since aquatic fauna such as fish (~1200 species, Rainboth, 1996), molluscs, crustaceans and insects, are all highly dependent on this basin as a breeding ground (Davidson et al., 2006; Zalinge and Thuok, 1998).

Macroinvertebrates are a key component of freshwater ecosystems (Bogan, 2008; Palmer et al., 1997). In river systems, macroinvertebrate communities are differently organized and structured along environ-

[*] Corresponding author at: Laboratoire Evolution & Diversité Biologique, UMR 5174, Université Paul Sabatier – Toulouse III, 118 route de Narbonne, 31062 Toulouse cédex 4, France.
E-mail address: sorsim.ratha@gmail.com (R. Sor).

http://dx.doi.org/10.1016/j.limno.2017.04.001
Received 21 November 2016; Received in revised form 2 March 2017; Accepted 4 April 2017
Available online 10 April 2017

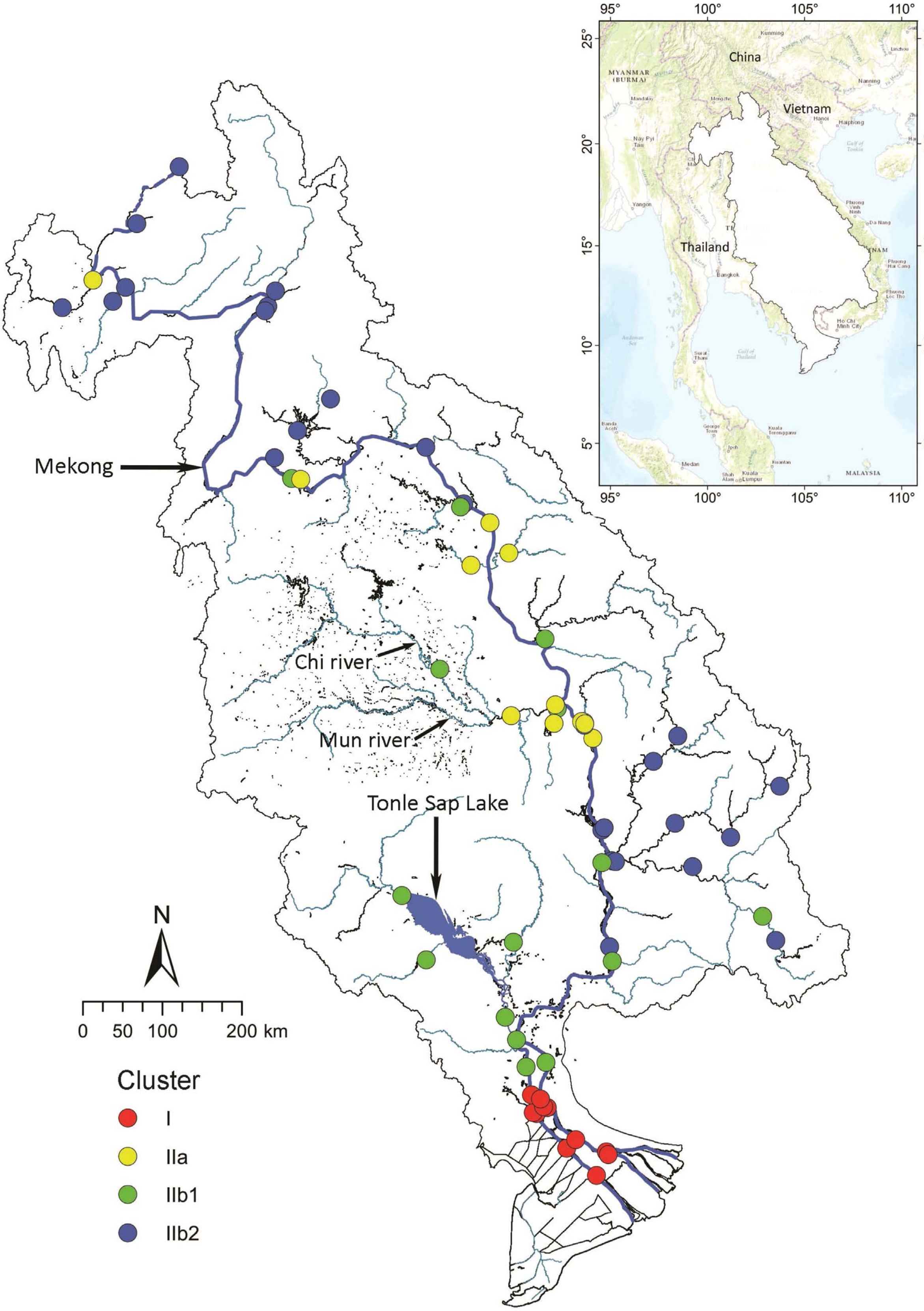


Fig. 1. Sampling sites and the four clusters, representing four macroinvertebrate assemblages, based on the cluster analysis.

mental gradients. Many studies have shown that in the upstream regions, the invertebrate communities are mainly characterized by a high abundance of insects (e.g. Ephemeroptera, Odonata and Trichoptera) (Arab et al., 2009; Collier and Lill, 2008; Jiang et al., 2013), whereas the downstream communities are characterized by a high abundance of molluscs, crustaceans and worms (Arscott et al., 2005;

Collier and Lill, 2008; Pérez-Quintero, 2011). This spatial variation in macroinvertebrate community composition is influenced by trophic variables, e.g. trophic level or food sources (Cai et al., 2012; Nicola et al., 2010); water chemistry, e.g. dissolved oxygen, water conductivity, pH (Al-Shami et al., 2013; Heino, 2009; Kumar and Khan, 2013) and physical river conditions, e.g. river depth, river width, size of catchment area and characteristics of land cover (Allan, 2004; Beisel et al., 1998; Chadwick et al., 2006; Cortes et al., 2013).

The Lower Mekong Basin (LMB) is characterized by a long and large floodplain (Eastham et al., 2008) and is known for its high biodiversity (Sodhi et al., 2004). However, the knowledge of macroinvertebrates in the LMB is limited. Given that this river basin is being impacted by various anthropogenic disturbances such as agricultural activities, aquaculture, urbanization and mining (Köhler et al., 2012; Nhan et al., 2007; Sodhi et al., 2004), there is an urgent need to understand the patterns of spatial organization and community structure of macroinvertebrates in this basin and their relation to environmental factors. Up to date, only a few studies have been published on the basin, e.g. community structure and composition of littoral invertebrates in the Mekong delta (Wilby et al., 2006), the diversity and distribution of crustaceans and molluscs in the Indo-Burma region (Cumberlidge et al., 2011; Köhler et al., 2012), and the assessment of water quality using physicochemical variables and benthic macroinvertebrates in the Chi river in Thailand (Kudthalang and Thanee, 2010). Yet, no attempt has been made to examine the spatial patterns of macroinvertebrate assemblages and their relation to key environmental variables at a large scale.

The objectives of the present study were i) to investigate patterns of spatial variation in macroinvertebrate assemblages in the Lower Mekong Basin, ii) to analyse the variability of macroinvertebrate composition among the assemblages, and to determine their key indicator taxa (the most representative taxa) for particular assemblages, and iii) to identify the important environmental variables that are associated with the macroinvertebrate assemblages in the basin. We expected that physical conditions of habitats, compared to other measured variables, have a strong correlation to the composition and diversity of macroinvertebrate assemblages.

2. Materials and methods

2.1. Study area

The Mekong River Basin (MRB) is divided into the Upper Mekong Basin (UMB) and the Lower Mekong Basin (LMB). The UMB on the Tibetan plateau in China is composed of narrow, deep gorges and small, short tributaries, whereas the LMB stretches from Yunnan province in South China to the delta in Vietnam and covers approximately 70% of the total length of the MRB (Eastham et al., 2008). The LMB consists of a large floodplain and long, broad tributaries and it drains more than 76% of the Mekong basin. The climate of the LMB is dominated by a tropical monsoon rainfall system, which is characterized by a dry (November − April) and a wet (May − October) season generated by the northeast monsoon and the southwest monsoon, respectively. The most intensive rainfall falls from July to September, while the lowest precipitation is observed between January and April (Adamson et al., 2009). The annual rainfall of the LMB varies from 1000–1600 mm in the driest regions to 2000–3000 mm in the wettest regions (Hoanh et al., 2003). A higher precipitation is found in the eastern mountainous regions of Laos and in northeast Thailand (Eastham et al., 2008).

The largest floodplain water body of the LMB is the Tonle Sap Lake (TSL) in Cambodia (Adamson et al., 2009), which is the largest freshwater lake in Southeast Asia (Sarkkula et al., 2003). The TSL is connected to the Mekong through the Tonle Sap River, and thus creating an exceptional hydrological cycle. In the wet season, the TSL receives an excess water from the Mekong River and expands its surface area from 2500 km^2 to 15,000 km^2. In the dry season when the rain ceases and water levels drop in the Mekong, a reverse flow occurs; the drained water from the TSL flows to the Mekong delta (Arias et al., 2011). The Mekong delta is characterized by a number of man-made canals, which are mostly used for domestic and agricultural activities (Kummu et al., 2008).

2.2. Data collection and processing

Benthic macroinvertebrates were sampled at 60 sampling sites along the main channel of the LMB and its tributaries by the Mekong River Commission (MRC) (Fig. 1). This sampling was carried out once a year in March during the dry season from 2004 to 2008. At each sampling site, macroinvertebrates were sampled from three locations in the benthic zone: near the left and right banks, and in the middle of the rivers. At each location, a minimum of three samples (where inter-sample variability is low, e.g. tributaries) and a maximum of five samples (where inter-sample variability is higher, e.g. the main channel and the delta) were collected using a Petersen grab sampler which has a sampling area of 0.025 m^2. With the grab sampler, four sub-samples were taken and pooled to give a single sampling unit covering a total area of 0.1 m^2. In total, between nine (3 samples × 3 locations) and fifteen (5 samples × 3 locations) pooled samples were collected at each sampling site. Each pooled sample was rinsed using a sieve (300 μm mesh size). In the field, the samples were sorted and then preserved by adding 10% formaldehyde to obtain a final concentration of about 5%. In the laboratory, they were identified to the lowest taxonomic level possible and counted using a compound microscope (40–1200 magnification) or a dissecting microscope (16–56 magnification). Macroinvertebrate abundance data per sampling unit was averaged across all samples (between 9 and 15 samples) collected from each sampling site.

At the sampling sites, geographic coordinates and altitude were determined with a GPS (Garmin GPS 12XL). River width was measured in the field using a Newcon Optik LRB 7 × 50 laser rangefinder. Other physical-chemical variables were measured at the three locations where macroinvertebrates were sampled. River depth was measured using a line metre. With a handheld water quality probe (YSI 556MP5), water temperature, dissolved oxygen, pH and water conductivity were measured at the surface (0.1–0.5 m) and at a depth of 3.5 m or at a maximum depth of the river (wherever less than 3.5 m) and then the average value was recorded for each location. Water transparency was measured with a Secchi disc by lowering it into the water and recording the depth at which it was no longer visible. The physical-chemical data of each sampling site was represented by the average value across the three sampling locations. The surface area and land cover data of watersheds drained at each sampling site were determined using a Geographic Information System (ArcGIS 10.4, ESRI). Geographic data (ArcGIS shapefiles) about the LMB (land cover types, river networks, basin boundaries and subcatchments derived from topographical maps) was provided by the MRC.

In total, 108 samples were collected from the 60 sampling sites. In 2008, 3 sampling sites were sampled further away from their original sampling coordinates, and thus we considered them as different sampling sites (see Supporting Information Appendix S1). Therefore, a total of 63 sampling sites were taken into account in the present study. Because of unequal sampling efforts (i.e. unequal and different number of samples at each site during the 5-year sampling period) and missing values of environmental variables, we used median values from the collected data to represent each site in our analysis, as suggested by McCluskey and Lalkhen (2007). More precisely, for the sites (49 sites) which were sampled only one time during the 5-year sampling period (see Appendix S1), the one sample collected from each site was used as the representative sample. The remaining 14 sites contained two samples (11 sites), three samples (2 sites) and four samples (1 site). From these 14 sites, the median values were used to represent the corresponding sites.

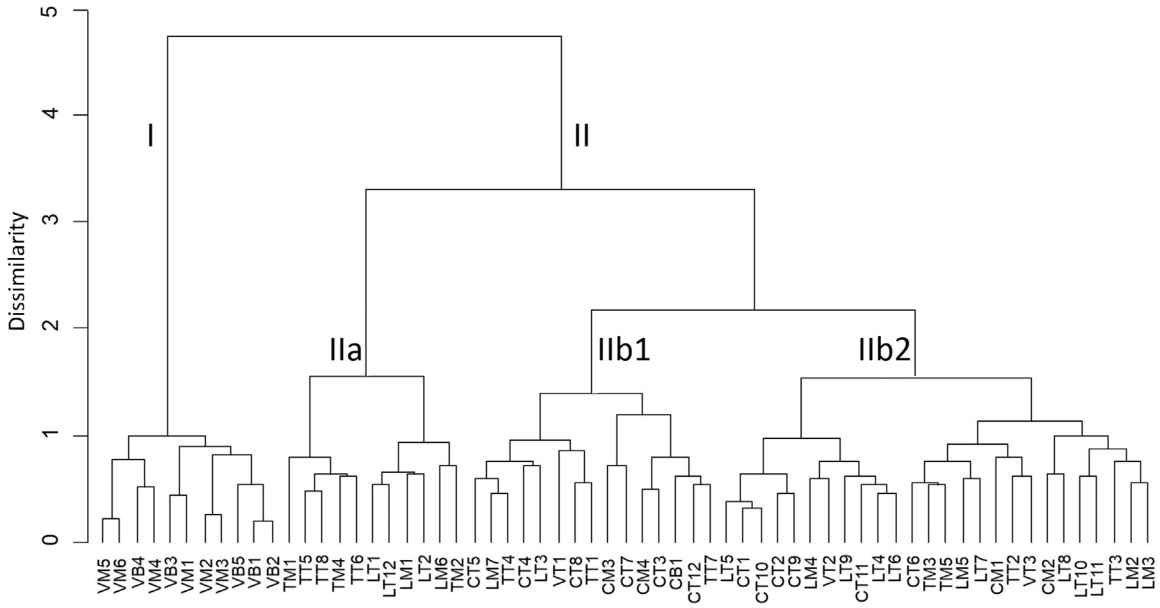


Fig. 2. Dendrogram showing sites belonging to the four clusters, representing four macroinvertebrate assemblages, based on the Bray-Curtis dissimilarity and Ward's hierarchical clustering method.

2.3. *Data analysis*

We used Ward's hierarchical clustering method to group the 63 sampling sites into different clusters. The Bray-Curtis dissimilarity distance (Legendre and Legendre, 2012) of macroinvertebrate samples between sites was calculated using the Hellinger transformation in the package "vegan" of R (Rao, 1995). Macroinvertebrate richness, abundance, Shannon-Wiener diversity were used to analyse the variability in macroinvertebrate assemblage composition among the different clusters. An "assemblage" of macroinvertebrates refers to different taxonomic groups (i.e. annelids, crustaceans, insects and molluscs) living in a given geographic area (e.g. a defined cluster) (Fauth et al., 1996). The indicator taxa of macroinvertebrate assemblages in the defined clusters were determined using the Indicator Value (IndVal, Dufrene and Legendre, 1997) with the package "labdsv" of R (Roberts, 2013). The Indicator Value of a taxon is an index ranging from 0 to 1, indicating the least to the most important taxa occurring in a group of sites. A value of 1 is obtained when every individual of the taxon is found only in the group and when it occurs in all sites of that group. A high number of taxa with significant Indicator Values may provide information on the habitat they prefer to share. Taxa having Indicator Values with a p-value ≤ 0.01 were retained as the most important taxa representing the assemblage in each cluster (consisting of a group of sites).

Significant differences in macroinvertebrate assemblage composition (i.e. richness, abundance, diversity) and in environmental conditions among different clusters were tested using a one-way ANOVA or a Kruskal-Wallis test. One-way ANOVA was applied when residuals of the models were normal (Shapiro-Wilk test, $p > 0.05$, and homoscedastic (Bartlett's test, $p > 0.05$)); otherwise, the non-parametric test (Kruskal-Wallis) was used. To evaluate the differences between clusters, multiple pair-wise comparisons were conducted using a Kruskal-Wallis test.

Linear Discriminant Analysis (LDA) was performed, using the package "ade4" of R (Chessel, 2006), to assess which measured environmental variables best accounted for the differences in macroinvertebrate assemblage composition among clusters grouped by the hierarchical clustering. Before performing the LDA, environmental variables were tested for multivariate homogeneity of within-group covariance (Borcard et al., 2011). The contribution of each variable to the discrimination between clusters was represented by the standardized factorial coefficient, projected as an arrow on the LDA plot. Cross-validation was used to evaluate the predictive performance of the LDA model. The overall quality of the model and the reliability of the prediction for the member of each cluster were evaluated by Cohen's Kappa Statistic (Kappa). All statistical analyses were performed in R (R Core Team, 2013).

3. Results

3.1. *Macroinvertebrate diversity*

Two hundred and ninety-nine macroinvertebrate taxa accounting for a total of 21,810 individuals were identified; of which, 131 taxa (44%) were insects, 98 (33%) were molluscs, 38 (13%) were crustaceans and 32 (10%) were annelids. The most commonly found insect orders were Diptera (37 taxa), Ephemeroptera (32), Odonata (22) and Trichoptera (20). For molluscs, most taxa belonged to the order Unionida (18), Veneroida (15) and Caenogastropoda (50); whereas for crustaceans, most taxa belonged to the order Decapoda (18) and Amphipoda (9). Annelids were mainly represented by the order Haplotaxida (15). Shannon diversity of macroinvertebrates at each site ranged from 0.8 to 3.3.

3.2. *Macroinvertebrate assemblages and indicator taxa*

Based on the Bray-Curtis dissimilarity and the hierarchical analysis, the 63 sites were grouped into four clusters (Fig. 2). Cluster I consisted of 11 sites located in the Mekong delta in Vietnam; cluster IIa, of 11 upstream sites in Laos and Thailand; cluster IIb1, of 15 sites mostly in middle part in Cambodia, a few sites in Laos and Thailand, and 1 site in Vietnam; and cluster IIb2, of 26 sites mostly located in the tributaries of the LMB (Fig. 1). The number of sampling sites and environmental variables characterizing each cluster are shown in Table 1.

The four clusters, which were later in the text considered as four different assemblages, had a significant difference in macroinvertebrate richness (Kruskal-Wallis $H = 27.7$, $p < 0.001$), abundance (Kruskal-Wallis $H = 23.7$, $p < 0.001$) and Shannon diversity (Kruskal-Wallis $H = 11.7$, $p < 0.01$) (Fig. 3, Table 1). The highest macroinvertebrate richness, abundance and diversity were found for assemblage I, followed by assemblage IIa and IIb1. The lowest macroinvertebrate richness, abundance and diversity were found for assemblage IIb2 (Fig. 3A and B, Table 1). Molluscs represented the highest proportion of

Table 1
Mean value (and standard deviation) for richness, abundance and Shannon diversity of macroinvertebrate assemblage and for environmental variables in each cluster.

	Cluster [n]							
Variable (unit)	I [11]		IIa [11]		IIb1 [15]		IIb2 [26]	
Assemblage composition								
Richness*	54	(13)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	23	(11)	18	(7)	16	(9)
Abundance*	955	(526)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	251	(184)	233	(204)	193	(286)
Diversity*	2.7	(0.7)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	2.1	(0.5)	2.0	(0.5)	1.9	(0.4)
Physical-chemical								
Altitude (m)*	6.6	(1.8)$^{-\mathrm{IIa}, -\mathrm{IIb2}}$	136	(77)	63	(67)$^{-\mathrm{IIb2}}$	207	(1539)
River width (m)*	1057	(468)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	413	(372)	349	(412)	339	(375)
River depth (m)*	11.5	(3.5)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	5.0	(3.3)$^{-\mathrm{IIb2}}$	5.0	(4.3)$^{-\mathrm{IIb2}}$	2.5	(1.7)
Secchi depth (m)*	0.6	(0.2)$^{-\mathrm{IIb2}}$	0.8	(0.4)	0.7	(0.5)	1.0	(0.7)
WT (°C)*	29.6	(0.5)$^{-\mathrm{IIa}, -\mathrm{IIb2}}$	26.0	(2.1)$^{-\mathrm{IIb1}}$	28.9	(1.6)$^{-\mathrm{IIb2}}$	26.4	(3.7)
DO (mg/L)*	6.2	(1.2)$^{-\mathrm{IIa}, -\mathrm{IIb2}}$	7.9	(0.5)$^{-\mathrm{IIb1}}$	6.1	(1.7)$^{-\mathrm{IIb2}}$	7.7	(0.7)
pH	7.8	(0.4)	7.6	(0.6)	7.5	(0.3)	7.6	(0.5)
WC (mS/m)*	17.8	(1.4)$^{-\mathrm{IIa}, -\mathrm{IIb2}}$	22.8	(6.0)$^{-\mathrm{IIb1}}$	13.5	(6.4)$^{-\mathrm{IIb2}}$	14.2	(9.8)
SAW (km^2)*	764,797	(4714)$^{-\mathrm{IIa}, -\mathrm{IIb1,2}}$	180,454	(202,943)	187,351	(276,932)	123,341	(196,952)
Land cover (%)								
Agricultural land*	25.77	(0.29)$^{-\mathrm{IIb2}}$	24.56	(26.7)	28.14	(24.82)$^{-\mathrm{IIb2}}$	11.68	(12.71)
Bamboo*	0.47	(0)	0.17	(0.22)$^{-\mathrm{IIb2}}$	0.61	(1.37)	2.14	(3.6)
Crops	5.48	(0.03)	5.53	(3.71)	3.97	(3.09)$^{-\mathrm{IIb2}}$	8.59	(7.62)
Deciduous forests	10.02	(0.13)	15.1	(26.54)	20.55	(18.28)$^{-\mathrm{IIb2}}$	9.01	(15.44)
Evergreen forests*	14.07	(0.06)	10.03	(5.5)$^{-\mathrm{IIb2}}$	14.07	(10.52)	20.35	(15.62)
Glacier	0.1	(0)	0.11	(0.14)	0.04	(0.08)	0.08	(0.14)
Grassland	11.6	(0.06)	11.55	(13.89)	5.96	(7.32)	10.76	(13.22)
Inundated*	0.39	(0.01)$^{-\mathrm{IIa}, -\mathrm{IIb2}}$	0	$^{-\mathrm{IIb1}}$	0.5	(0.96)$^{-\mathrm{IIb2}}$	0	
Mix_evg.dec	8.99	(0.05)	12.82	(11.64)	9.68	(6.49)	8.79	(7.9)
Plantations*	0.17	(0)	0.03	(0.07)	0.2	(0.35)	0.17	(0.3)
Regrowth*	0.88	(0.01)	0.31	(0.3)$^{-\mathrm{IIb2}}$	0.98	(1.1)	1.25	(1.58)
Rocks*	0.6	(0)	1.46	(3.34)$^{-\mathrm{IIb1,2}}$	0.23	(0.31)	0.4	(0.54)
Urban areas*	0.08	(0)	0.51	(1.49)$^{-\mathrm{IIb1,2}}$	0.07	(0.06)	0.07	(0.11)
Water surface*	1.18	(0.01)$^{-\mathrm{IIb2}}$	0.82	(1.57)	1.18	(1.26) $^{-\mathrm{IIb2}}$	0.42	(0.87)
Wetland*	0.07	(0.01)	0.02	(0.02)	0.09	(0.18) $^{-\mathrm{IIb2}}$	0.01	(0.02)
Wood- & shrub-land*	17.23	(0.07)	14.05	(10.68)$^{-\mathrm{IIb2}}$	12.5	(7.91) $^{-\mathrm{IIb2}}$	24.38	(15.76)

WT: water temperature, DO: dissolved oxygen, WC: water conductivity, SAW: the surface area of watersheds, Mix_evg. dec: mixed evergreen and deciduous forests. The number of samples [n] in each cluster is indicated between square brackets. * indicates ANOVA and Kruskal-Wallis Test for significant differences among clusters at $p < 0.05$. Superscripts (IIa, IIb1, IIb2) indicate significant pair-wise comparisons between the corresponding cluster (each column) and superscript-labeled clusters (i.e. IIa, IIb1, IIb2) at $p < 0.05$.

richness and abundance for assemblage I and IIb1. For assemblage IIa and IIb2, insects represented the highest proportion, followed by molluscs. Crustaceans and annelids made up the lowest proportion of richness and abundance for assemblage IIa, IIb1 and IIb2, but they represented a higher proportion compared to insects for assemblage I (Fig. 3C and D).

The four assemblages were represented by a different number of indicator taxa: 53, 14, 2 and 2 taxa, for assemblage I, IIa, IIb1 and IIb2, respectively. Molluscs, crustaceans and annelids represented a relatively high number of indicator taxa compared to insects for assemblage I. The assemblage IIa was mostly represented by insects, while assemblages IIb1 and IIb2 were represented by two different indicator taxa. The detailed information on indicator taxa for each assemblage is provided in Table 2.

3.3. Relationship between macroinvertebrate assemblages and environmental variables

Most of the environmental variables showed significant differences between the four macroinvertebrate assemblages. The detailed information on the environmental differences between the assemblages is provided in Table 1. The results of the LDA used to discriminate the macroinvertebrate assemblages based on the physical-chemical and land cover types are shown in Fig. 4. The factorial axes 1, 2 and 3 respectively explained 65.7%, 23.1% and 11.2% of the total variance of the data. The global performance of the LDA was high (Kappa = 0.86). The predictive reliability (Kappa) for assemblage I, IIa, IIb1 and IIb2 was 0.90, 0.84, 0.80 and 0.90, respectively. Along axis 1, assemblage I was situated opposite to assemblage IIa and IIb2. Assemblage I was positively correlated with the surface area of watershed, river depth, river width and water temperature, but negatively associated with altitude and dissolved oxygen. Whereas assemblage IIa was positively correlated with water conductivity and urban areas, and assemblage IIb2 was positively linked to high altitude, dissolved oxygen, Secchi depth, wood-/shrub-land and evergreen forests. Based on axis 1 and 3, assemblage IIb1 was positively linked to inundated, wetland and agricultural areas (Fig. 4A–D).

4. Discussion

4.1. Spatial variation in macroinvertebrate assemblages and habitat characteristics

Macroinvertebrate assemblages can vary substantially according to habitat characteristics, particularly with large variations in habitat types and resources (Costa and Melo, 2008; Ilg et al., 2008; Vaughn and Hakenkamp, 2001). The combination of environmental factors such as water velocity, river depth, organic material load, watershed and substratum structure, together with the high amount of food resources and different water quality conditions found in the LMB (Chea et al., 2016), are responsible for determining macroinvertebrate assemblages (Al-Shami et al., 2013; Beisel et al., 1998; Lamouroux et al., 2014; Lorenz and Feld, 2013; Pan et al., 2014). Moreover, macroinvertebrate assemblages are also structured by different land cover characteristics, e.g. agricultural land, wood-/shrub-land and urban areas (Allan, 2004; Cortes et al., 2011, 2013). Therefore, these physical-chemical variables and land cover types are likely the main factors determining the spatial organization and composition of macroinvertebrate assemblages in this

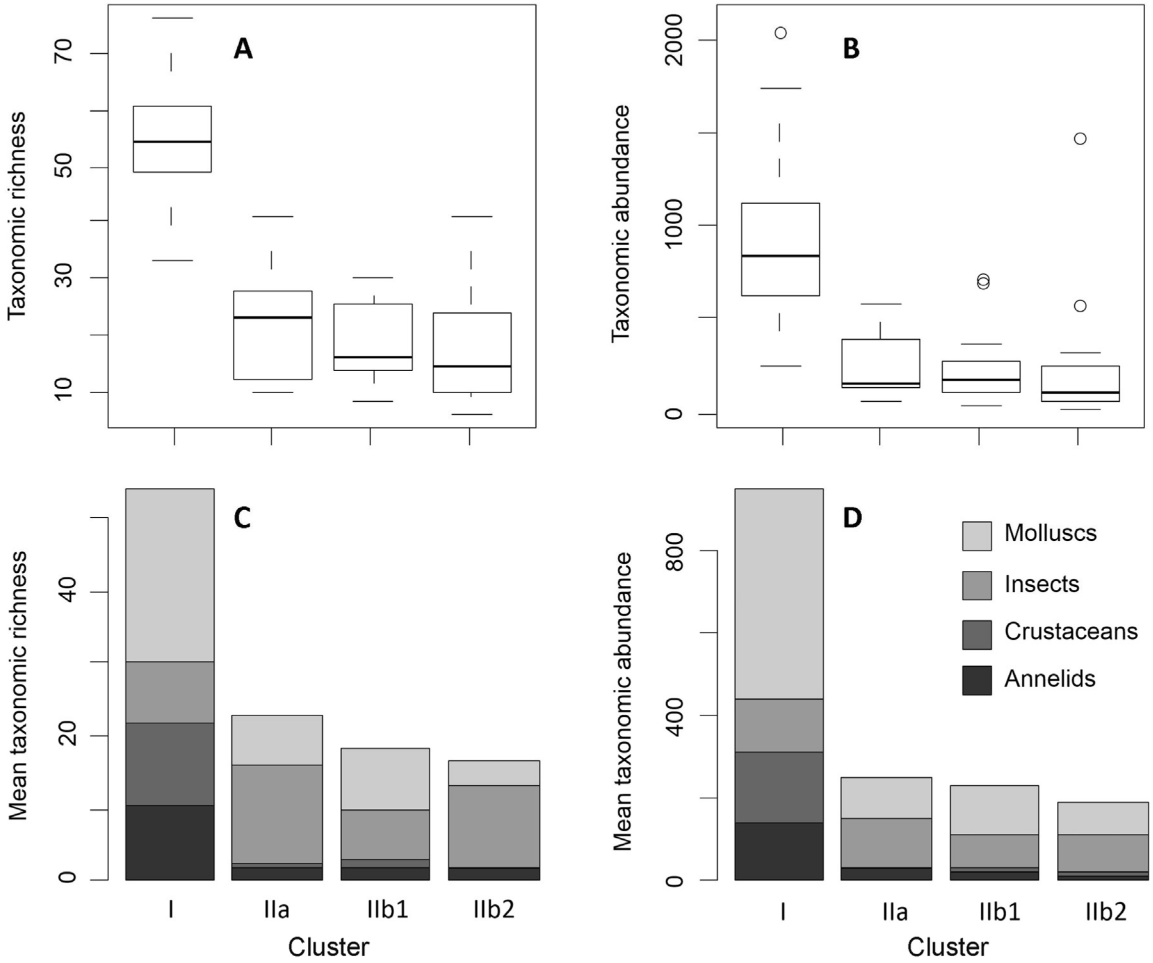


Fig. 3. Box and whisker plots of richness (A) and abundance (B) of macroinvertebrate assemblage in each cluster and its proportion of mean richness (C) and abundance (D) consisting of different taxonomic groups of macroinvertebrates.

large floodplain river.

The spatial organization of macroinvertebrate assemblages is mainly related to the relationship between habitat conditions and life-history traits at different spatial scales or along the longitudinal gradients of rivers (Collier and Lill, 2008; Tonkin et al., 2012). In this regard, key physical-chemical variables and land cover types can act as environmental filters, which allow the most suited taxonomic groups of macroinvertebrates to be present in each assemblage (Menezes et al., 2010). We found a higher insect richness and abundance for assemblage IIb2, which is mostly related to the tributaries and some sites of the main channel at the upstream parts. This assemblage is represented only by one insect and one annelid indicator taxon, which were widely distributed in the assemblage. This result suggests that natural stress in the mountainous and tributary system is probably too high for some taxa to occur (Feld and Hering, 2007). The fact that this assemblage was dominated by insects agrees with the expectation that insect orders mainly inhabit mountainous and shaded habitat regions (Ferrington, 2008; Heino, 2009; Suhaila and Che Salmah, 2014). The presence of evergreen forests, bamboo, crops and woody debris may provide food resources and egg deposition sites for many insect taxa (Medhurst et al., 2010; Sweeney, 1993). Clear water and a high level of dissolved oxygen found in high altitude streams characterize the tributaries, which are preferred by insect taxa (Collier and Lill, 2008; Dobson et al., 2002; Królak and Korycińska, 2008). Moreover, sampling sites in this assemblage are comparable to the Holarctic region where Diptera are more abundant (Ferrington, 2008), and are characterized by the increasing latitude where most Ephemeroptera occur (Pearson and Boyero, 2009).

Macroinvertebrate assemblage IIa was mainly related to the upstream sites situated along the main channel and some tributaries. In this assemblage, insects were also the dominant group in terms of richness, abundance and indicator taxa (7 out of 14 taxa, Table 2). However, annelid and mollusc indicator taxa were also found in this assemblage. A high water conductivity (22.8 ± 6 mS/m), resulting from anthropogenic disturbance (e.g. runoff from the surrounding agricultural areas and the discharge of urban wastewaters), could be one of the factors responsible for the community composition found, as it has been demonstrated that conductivity has a strong impact on macroinvertebrate diversity (Kumar and Khan, 2013; Lods-Crozet et al., 2001). The high contribution of insects to this assemblage could be explained by a high level of dissolved oxygen present at some sites (Table 1), as indicated in previous studies (Dobson et al., 2002; Królak and Korycińska, 2008), while disturbance-tolerant taxa of Oligochaeta and Gastropoda may be well adapted at sites having a high water conductivity (Feld and Hering, 2007; Wang et al., 2012). Moreover, the main channel generally provides a higher variation in habitats and a higher nutrient and sediment load, which are also favourable for annelids and molluscs (Haag and Warren-Jr., 1998).

In-between the upstream and downstream parts of the river basin, the macroinvertebrate assemblage was dominated by molluscs, followed by insects. Characteristics of sites in the mid-reaches such as rock and deciduous forests play an important role in supplying organic matter, which is preferred by snails (grazers) and insect collectors and shredders (Thorp and Delong, 1994). Many sites belonging to this assemblage that are located around the Tonle Sap Lake, Cambodia, are characterized by inundated, agricultural land and wetland (Fig. 4C and

Table 2
List of indicator taxa (and their indicator values, IndVal) of macroinvertebrate assemblage in each cluster.

Cluster I	IndVal	p-value
Annelid		
Aeolosoma bengalense	0.52	0.01
Aulodrilus prothecatus	0.67	0.005
Chaetogaster langi	0.85	0.005
Chaetogaster limnaei limnaei	0.6	0.005
Dero pectinata	0.67	0.005
Dero sp.	0.74	0.005
Dero sp.1	1	0.005
Dero sp.2	0.95	0.005
Lumbriculidae sp.	0.6	0.005
Namalycastis longicirris	0.9	0.005
Nectopsyche sp.	0.67	0.005
Orbinia johnsoni	0.52	0.01
Polydora sp.	0.67	0.005
Crustacean		
Corophium minutum	0. 8	0.005
Corophium sp.	0.67	0.005
Cyathura carinata	0.74	0.005
Cyathura truncata	0.57	0.005
Decapoda sp.	0.91	0.005
Eohaustorius sp.	0.6	0.005
Eohaustorius tandeensis	0.67	0.005
Gammarus sp.	0.6	0.005
Grandidierella lignorum	0.78	0.005
Grandidierella vietnamica	1	0.005
Hyale hawaiensis	0.67	0.005
Hyale sp.	0.85	0.005
Kamaka sp.	0.6	0.005
Macrobrachium equidens	0.6	0.005
Melita sp.	0.82	0.005
Palaemon curvirostris	0.6	0.005
Mollusc		
Afropisidium clarkeanum	0.73	0.005
Angulyagra polyzonata	0.6	0.005
Angulyagra sp.	0.9	0.005
Bithynia siamensis	0.67	0.005
Corbicula baudoni	0.87	0.005
Corbicula bocourti	0.74	0.005
Corbicula leviuscula	0.97	0.005
Corbicula moreletiana	0.86	0.005
Corbicula sp.	0.95	0.005
Gastropoda sp.	0.74	0.005
Hyriopsis bialatus	0.64	0.005
Limnoperna siamensis	0.99	0.005
Limnoperna sp.	0.95	0.005
Lymnaea viridis	0.94	0.005
Mekongia swainsoni swainsoni	0.67	0.005
Sinomytilus harmandi	0.9	0.005
Stenothyra annandalei	0.6	0.005
Stenothyra glabrata	0.85	0.005
Trochotaia trochoides	0.52	0.005
Insect		
Arigomphus sp.	0.67	0.005
Cricotopus sp.	1	0.005
Clinotanypus sp.	0.52	0.005
Monocorophium sp.	0.91	0.005
Sigara sp.	0.6	0.005
Cluster IIa	IndVal	p-value
Annelid		
Oligochaeta sp.	0.99	0.005
Polychaeta sp.1	0.6	0.005
Mollusc		
Corbicula sp.	0.88	0.005
Hubendickia sp.	0.6	0.01
Kareliania sp.	0.52	0.01
Scaphula sp.	0.52	0.01
Stenothyra sp.	0.6	0.005
Insect		
Anagenesia sp.	0.67	0.005
Caenoculis sp.	0.52	0.01
Caenodes sp.	0.74	0.005
Choropterpes sp.	0.51	0.005
Dipseudopsis sp.	0.69	0.005
Heterocloeon sp.	0.52	0.01
Micronecta sp.	0.6	0.005
Cluster IIb1	IndVal	p-value
Mollusc		
Filopaludina filopaludina filosa	0.45	0.025
Insect		
Pentagenia sp.	0.62	0.01
Cluster IIb2	IndVal	p-value
Annelid		
Naididae sp.	0.76	0.005
Insect		
Gomphidae sp.	0.56	0.01

D, Table 1). These site specific features could promote the richness and abundance of molluscs found at these sites. Moreover, it is noteworthy that the areas around the Tonle Sap Lake support a high mollusc production (Ngor et al., 2016). Therefore, it is not surprising that molluscs made up the largest proportion of the macroinvertebrate assemblage and that snail species were put forward as indicator species for this assemblage. The tributary sites that are located upstream, characterized by shaded areas and surrounded by large aquatic plants, are likely responsible for the second highest proportion of insect richness and abundance and the presence of an insect indicator taxon.

In the LMB, downstream sites were mainly associated with a large surface area of watersheds, large floodplains, wide and deep rivers. These conditions promote a high nutrient and sediment load from the upstream parts (Blair et al., 2004), which can support many species and a high abundance of macroinvertebrates (Cai et al., 2012; Nicola et al., 2010). Moreover, the high temperature related to the downstream sites may also enhance the richness and abundance of molluscs (Vaughn and Hakenkamp, 2001). High nutrient input and high temperature due to sunlight observed at the downstream delta provide optimal conditions for the phytoplankton community to reach a high abundance (Hecky and Kilham, 1988; Statzner and Higler, 1985; Vannote et al., 1980), which in turn enhance the richness and abundance of phytoplankton feeders such as molluscs (Vaughn and Hakenkamp, 2001). However, although the downstream assemblage was dominated by molluscs, annelids and crustaceans, dipteran insects (14 out of 33 occurring taxa) also made up a relatively high proportion. As a result, each taxonomic group in this assemblage is represented by many indicator taxa. This suggests that the downstream sites, the delta, provide good habitat conditions to support a diverse fauna. The high richness and abundance of all taxonomic groups and the high diversity in this assemblage might be explained 1) by the large surface area of the watershed and the large width and depth of the river, which may provide an optimal nutrient load and different mesohabitats and microhabitats, respectively (Al-Shami et al., 2013; Haag and Warren-Jr., 1998; Jacobsen et al., 1997; Mereta et al., 2012; Sedell et al., 1989), and 2) by a slow water flow, as reflected by the low change in altitude (6.6 ± 1.8 m) of the sampling sites, which may provide good habitat conditions for less-mobile taxa like molluscs, annelids and crustaceans (Castella et al., 1994; Haag and Warren-Jr., 1998).

The compositional and diversity differences found for the four macroinvertebrate assemblages were related to spatial environmental heterogeneity observed from the upstream to the downstream parts of the LMB (Dobson et al., 2002; Heino et al., 2005; Salmah et al., 2014; Wang et al., 2012). This can be explained by the proportion of the

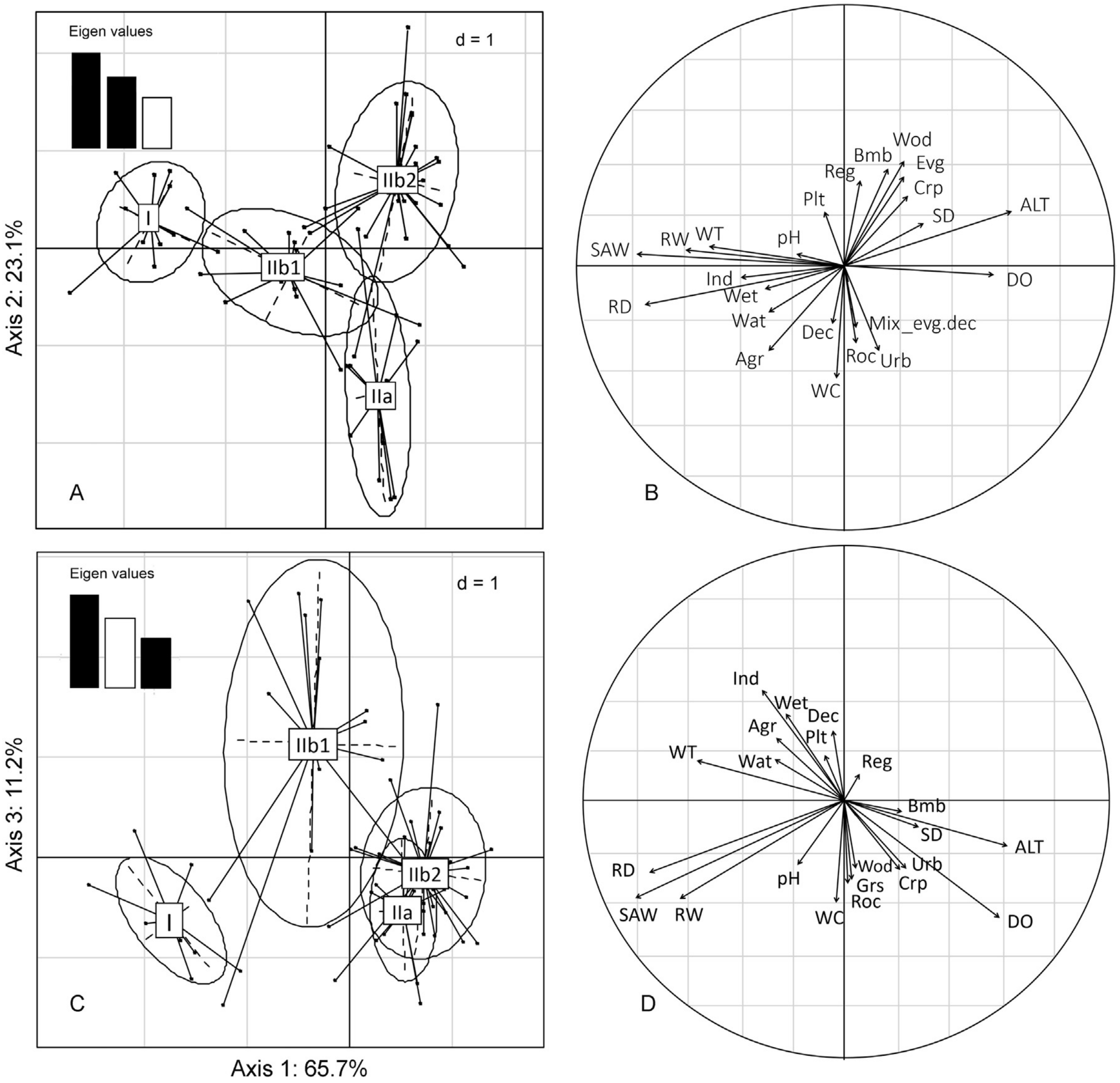


Fig. 4. Results from the LDA discriminating the four clusters (I, II, IIb1, IIb2), representing four macroinvertebrate assemblages, using Axes 1, 2 and 3 that explained the indicated percentage of the total variance in the data (A, C), and correlations of the environmental factors to the corresponding axes (B, D). ALT: altitude, RW: river width, RD: river depth, SD: Secchi depth, WT: water temperature, DO: dissolved oxygen, WC: water conductivity, SAW: the surface area of watersheds, Agr: agricultural land, Bmb: bamboos, Crp: crops, Dec: deciduous forests, Evg: evergreen forests, Grs: grassland, Ind: inundated, Mix_evg. dec: mixed evergreen and deciduous forests, Plt: plantations, Reg: regrowth, Roc: rocks, Urb: urban areas, Wat: water, Wet: wetland, Wod: wood- & shrub-land.

richness, abundance, diversity and the represented indicator taxa of the different taxonomic groups for each assemblage. Although assemblage IIa, IIb1 and IIb2 were characterized by a low and similar macroinvertebrate richness, abundance and diversity, each taxonomic group (i.e. annelids, crustaceans, molluscs and insects) contributed differently to the overall macroinvertebrate community composition for the particular assemblages. On the other hand, assemblage I had a high richness, abundance, diversity and a high number of indicator taxa. Therefore, these findings clearly suggested that the community composition of the four assemblages, particularly assemblage I, is distinctly organized and structured along the environmental gradients of the LMB.

4.2. Relationship between assemblages and key environmental variables

Site-specific characteristics, as mentioned above, are known to influence local community and composition of macroinvertebrates. At an assemblage level, connection in habitats and mesohabitats and variation in river morphology strongly affect community structure and composition from the upstream to the downstream parts of rivers (Mazão and Bispo, 2016; Thorp and Delong, 1994; Zilli and Marchese, 2011). Our study demonstrated that among the measured environmental variables, physical conditions and land cover types are the key factors determining macroinvertebrate assemblages along the LMB. These results support the previous finding that macroinvertebrate community structures in the Chishui river basin and in the river-connected lakes of the Yangtze river are strongly influenced and well predicted by physical habitat variables, e.g. water depth and altitude (Jiang et al., 2010; Pan et al., 2014). In other tropical/sub-tropical (Pearl and Qiangtang rivers, southern China) and temperate river basins (e.g. the Olo, Corgo, Pinhao and Tua rivers in northern Portugal), characteristics of land cover have been reported to better explain and predict benthic macroinvertebrate communities (Allan, 2004; Cortes et al., 2011, 2013).

In the LMB, the surface area of watershed, river depth and altitude are the most important variables discriminating the upstream assemblages (IIa and IIb2) from the downstream assemblage (I). This discrimination is clearly explained by the axis 1 of the LDA accounting for 65.7% of the total variance of the data and by a high predictive reliability for each particular assemblage. Because of the observed physical conditions, together with other related variables (e.g. water temperature, dissolved oxygen and land cover types), the assemblages were characterized by a very different community structure, composition and diversity.

The intermediate assemblage (IIb1) in-between was positively linked to three types of land cover: agricultural land, inundated and wetland. Along axis 3 of the LDA model, these land cover types are the key factors best explaining the macroinvertebrate composition of assemblage IIb1 (Fig. 4C and D). However, the proportion of inundated and wetland constituted a small percentage of the land cover, and together with agricultural land, they only explained a limited amount of the variance of the data for axis 3 of the LDA model (11.2%). As a result, the predictive performance for this assemblage (Kappa = 0.80) was not as high as the performance for the other assemblages. A broad range of values of physical-chemical variables (e.g. dissolved oxygen, river width and depth, the surface area of watershed and altitude) found at sites belonging to this assemblage could also explain the performance yielded. On the other hand, assemblage IIa was associated with two important factors (water conductivity and urban areas) according to the LDA axis 2, which explained a higher variance of the data (23.1%). Therefore, these two factors, together with other measured environmental variables, could correctly predict this assemblage with a higher reliability (Kappa = 0.85). The possible explanation for the association found is that many sites that belong to this assemblage are located along the tributaries (e.g. the Mun and Chi river basins in Thailand) and the main channel where cities were built. The tributaries are surrounded by intensified agriculture and the cities are exposed to a high level of anthropogenic disturbance (Dao et al., 2010; Kudthalang and Thanee, 2010). Therefore, the runoff from the surrounding agricultural areas and the discharge of urban wastewaters may cause the increase in water conductivity (Wetzel, 2001).

4.3. Recommendations for management of the LMB

Macroinvertebrate diversity is of great importance for ecosystem structure and functioning as it interacts with both biotic and abiotic factors. Crustaceans and insects are generally responsible for regulating decomposition, shredding detritus and bioturbation. Molluscs also contribute to bioturbation, sediment formation and filtering of water, while most annelids regulate decomposition and autotrophs although some of them also promote bioturbation and sediment formation (Palmer et al., 1997). The different processes enhanced by these taxa promote ecosystem functioning in the LMB, particularly in its lower reaches. Assemblage I, occupying the downstream part of the LMB, was characterized by a high macroinvertebrate diversity and was represented by a number of indicator taxa from each group of macroinvertebrates, which could be expected to increase ecosystem stability. These taxa are key components of the ecosystem and of the food web supporting a rich biodiversity of fish, which are a main source of proteins for local people. Thus, a change in the composition of one of these taxa may result in disproportionate or unexpected responses of other taxa in the LMB (Naeem, 1998), and consequently alter ecological processes (Covich et al., 1999) such as organic matter processing (Palmer et al., 1997).

For management purposes, indicator taxa identified in this study representing the ecological importance of each site and each community are of great significance. By using the results of this study such as the indicator taxa and the correlated environmental factors for each assemblage, habitat quality and suitability for the most vulnerable native taxa could be more conveniently monitored, which could provide an advantage for decision-making concerning conservation and management of these keystone taxa. Moreover, this study can serve as a baseline for future research in a biodiversity hotspot region that has not been investigated intensively.

5. Conclusion

The macroinvertebrate assemblages in the Lower Mekong Basin (LMB) were characterized by an increasing richness, abundance and diversity from the upstream to the downstream sites. Sites located in the upper tributaries (cluster IIb2) and upper main channel (cluster IIa) were characterized by a high altitude, high levels of dissolved oxygen, a high water conductivity, large fractions of wood-/shrub-land and evergreen forests. Macroinvertebrate assemblages in these two clusters were dominated by insects. In the Mekong delta (cluster I), the macroinvertebrate assemblage largely consisted of a great number of molluscs and an average number of annelids, crustaceans and dipteran insects. This assemblage could be discriminated based on a large surface area of watershed, the river depth and width and a high water temperature. The assemblage (cluster IIb1) found in-between the tributaries and the upper main channel and the delta region was associated with sites characterized by inundated, wetland and agricultural land and was represented most by molluscs. Overall, our study found that the number of macroinvertebrate taxa and the diversity remain relatively high across the basin, especially in the delta region which can be considered a hotspot zone for biodiversity. Therefore, the LMB deserves a particular attention for conservation.

Acknowledgements

The first author is grateful to the Erasmus Mundus LOTUS project (European Commission) and the Forum Belmont TLSCC project for providing the opportunity to do this work. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. We would like to sincerely thank the Mekong River Commission who provided the database for our analyses. Many thanks go to L. Tudesque for his help in determining the surface area of watersheds and the percentage of land cover and K.K.Y. Wong for his editorial comments on this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.limno.2017.04.001.

References

Adamson, P.T., Rutherfurd, I.D., Peel, M.C., Conlan, I.A., 2009. The Hydrology of the Mekong River, 1st ed. Elsevier Inc, The Mekong.

Al-Shami, S.A., Heino, J., Che Salmah, M.R., Abu Hassan, A., Suhaila, A.H., Madrus, M.R., 2013. Drivers of beta diversity of macroinvertebrate communities in tropical forest streams. Freshw. Biol. 58, 1126–1137.

Allan, D., 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. Annu. Rev. Ecol. Evol. Syst. 35, 257–284.

Arab, A., Lek, S., Lounaci, A., Park, Y.S., 2009. Spatial and temporal patterns of benthic invertebrate communities in an intermittent river (North Africa). Ann. Limnol. − Int. J. Limnol. 40, 317–327.

Arias, M.E., Cochrane, T.A., Caruso, B., Killeen, T., Kummu, M., 2011. A landscape approach to assess impacts of hydrological changes to vegetation communities of the Tonle Sap Floodplain. In: Engineering: Conference Contributions. Canterbury. pp. 3018–3025.

Arscott, D.B., Tockner, K., Ward, J.V., 2005. Lateral organization of aquatic invertebrates along the corridor of a braided floodplain river. J. North Am. Benthol. Soc. 24, 934–954.

Beisel, J.N., Usseglio-Polatera, P., Thomas, S., Moreteau, J.C., 1998. Stream community structure in relation to spatial variation: the influence of mesohabitat characteristics. Hydrobiologia 389, 73–88.

Blair, N.E., Leithold, E.L., Aller, R.C., 2004. From bedrock to burial: the evolution of particulate organic carbon across coupled watershed-continental margin systems. Mar. Chem. 92, 141–156.

Bogan, A.E., 2008. Global diversity of freshwater mussels (Mollusca, Bivalvia) in freshwater. Hydrobiologia 595, 139–147.

Borcard, D., Gillet, F., Legendre, P., 2011. Numerical Ecology with R. Springer Science & Business Media, New York.

Cai, Y., Gong, Z., Qin, B., 2012. Benthic macroinvertebrate community structure in Lake Taihu, China: effects of trophic status, wind-induced disturbance and habitat

complexity. J. Great Lakes Res. 38, 39–48.
Castella, E., Speight, M., Obrdlik, P., Schneider, E., Lavery, T., 1994. A methodological approach to the use of terrestrial invertebrates for the assessment of alluvial wetlands. Wetl. Ecol. Manage. 3, 17–36.
Chadwick, M.A., Dobberfuhl, D.R., Benke, A.C., Huryn, A.D., Suberkropp, K., Thiele, J.E., 2006. Urbanization affects stream ecosystem function by altering hydrology, chemistry, and biotic richness. Ecol. Appl. 16, 1796–1807.
Chea, R., Grenouillet, G., Lek, S., 2016. Evidence of water quality degradation in lower mekong basin revealed by self organizing map. PLoS One 11, e0145527.
Chessel, D., 2006. The ade4 Package .
Collier, K.J., Lill, A., 2008. Spatial patterns in the composition of shallow-water macroinvertebrate communities of a large New Zealand river. N. Z. J. Mar. Freshw. Res. 42, 129–141.
Cortes, R., Varandas, S., Teixeira, A., Hughes, S., Magalhaes, M., Barquín, J., Álvarez-Cabria, M., Fernández, D., 2011. Effects of landscape metrics and land use variables on macroinvertebrate communities and habitat characteristics. Limnetica 30, 347–362.
Cortes, R.M.V., Hughes, S.J., Pereira, V.R., Varandas, S.D.G.P., 2013. Tools for bioindicator assessment in rivers: the importance of spatial scale, land use patterns and biotic integration. Ecol. Indic. 34, 460–477.
Costa, S.S., Melo, A.S., 2008. Beta diversity in stream macroinvertebrate assemblages: among-site and among-microhabitat components. Hydrobiologia 598, 131–138.
Covich, A.P., Palmer, M.A., Crowl, T.A., 1999. The role in of species invertebrate freshwater ecosystems − zoobenthic species influence energy flows and nutrient cycling. Bioscience 49, 119–127.
Cumberlidge, N., Ng, P.K.L., Yeo, D.C.J., 2011. Freshwater crabs of the Indo-Burma hotspot: diversity, distribution, and conservation. In: Allen, D.J., Darwall, W.R.T. (Eds.), The Status and Distribution of Freshwater Crabs. IUCN, Gland and Cambridge, pp. 102–113.
Dao, H., Kunpradid, T., Vongsambath, C., Do, T., Prum, S., 2010. Report on the 2008 biomonitoring survey of the lower Mekong River and selected tributaries. MRC Technical Paper No. 27. PDR, Vientiane, Lao.
Davidson, P.S., Kunpradid, T., Peerapornisal, Y., Nguyen, T.M.L., Pathoumthong, B., Vongsambath, C., Pham, A.D., 2006. Biomonitoring of the lower Mekong River and selected tributaries. MRC Technical Paper No. 13. PDR, Vientiane, Lao.
Dobson, M., Magana, A.E.M., Mathooko, J.M., Ndegwa, F.K., 2002. Detritivores in Kenyan highland streams: more evidence for the paucity of shredders in the tropics? Freshw. Biol. 47, 909–919.
Dufrene, M., Legendre, P., 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67, 345–366.
Eastham, J., Mpelasoka, F., Ticehurst, C., Dyce, P., Ali, R., Kirby, M., 2008. Mekong River Basin Water Resources Assessment: Impacts of Climate Change. Water for a Healthy Country National Research Flagship, CSIRO.
Fauth, J.E., Bernardo, J., Camara, M., Resetarits Jr., W.J., Buskirk, J., Van, McCollum, S.A., 1996. Simplifying the jargon of community ecology: a conceptual approach. Am. Nat. 147, 282.
Feld, C.K., Hering, D., 2007. Community structure or function: effects of environmental stress on benthic macroinvertebrates at different spatial scales. Freshw. Biol. 52, 1380–1399.
Ferrington, L.C., 2008. Global diversity of non-biting midges (Chironomidae; Insecta-Diptera) in freshwater. Hydrobiologia 595, 447–455.
Haag, W.R., Warren-Jr, M.L., 1998. Role of ecological factors and reproductive strategies in structuring freshwater mussel communities. Can. J. Fish. Aquat. Sci. 55, 297–306.
Hecky, R.E., Kilham, P., 1988. Nutrient limitation of phytoplankton in freshwater and marine environments: a review of recent evidence on the effects of enrichment. Limnol. Oceanogr. 33, 796–822.
Heino, J., Parviainen, J., Paavola, R., Jehle, M., Louhi, P., Muotka, T., 2005. Characterizing macroinvertebrate assemblage structure in relation to stream size and tributary position. Hydrobiologia 539, 121–130.
Heino, J., 2009. Biodiversity of aquatic insects: spatial gradients and environmental correlates of assemblage-level measures at large scales. Freshw. Rev. 2, 1–29.
Hoanh, C.T., Guttman, H., Droogers, P., Aerts, J., 2003. Water, Climate, Food, and Environment in the Mekong Basin in Southeast Asia − Final Report. International Water Management Institute, Mekong River Commission Secretariat.
Ilg, C., Foeckler, F., Deichner, O., Henle, K., 2008. Extreme flood events favour floodplain mollusc diversity. Hydrobiologia 621, 63–73.
Jacobsen, D., Schultz, R., Encalada, A., 1997. Structure and diversity of stream invertebrate assemblages: the influence of temperature with altitude and latitude. Freshw. Biol. 38, 247–261.
Jiang, X., Xing, J., Qiu, J.W., Wu, J.M., Wang, J.W., Xie, Z., 2010. Structure of macroinvertebrate communities in relation to environmental variables in a subtropical Asian river system. Int. Rev. Hydrobiol. 95, 42–57.
Jiang, X., Xie, Z., Chen, Y., 2013. Longitudinal patterns of macroinvertebrate communities in relation to environmental factors in a Tibetan-Plateau river system. Quat. Int. 304, 107–114.
Köhler, F., Seddon, M., Bogan, A.E., Tu, D., Van Sri-aroon, P., Allen, D., 2012. The status and distribution of freshwater molluscs of the Indo-Burma region. In: Allen, D., Smith, K., Darwall, W. (Eds.), The Status and Distribution of Freshwater Biodiversity in Indo-Burma. Gland, Cambridge, pp. 66–89.
Królak, E., Korycińska, M., 2008. Taxonomic composition of macroinvertebrates in the Liwiec River and its tributaries (Central and Eastern Poland) on the basis of chosen physical and chemical parameters of water and season. Pol. J. Environ. Stud. 17, 39–50.
Kudthalang, N., Thanee, N., 2010. The assessment of water quality in the upper part of the Chi basin using physicochemical variables and benthic macroinvertebrates. Suranaree J. Sci. Technol. 17, 165–176.
Kumar, P.S., Khan, A.B., 2013. The distribution and diversity of benthic macroinvertebrate fauna in Pondicherry mangroves, India. Aquat. Biosyst. 9, 1–18.
Kummu, M., Lu, X.X., Rasphone, A., Sarkkula, J., Koponen, J., 2008. Riverbank changes along the Mekong River: remote sensing detection in the Vientiane-Nong Khai area. Quat. Int. 186, 100–112.
Lamouroux, N., Dolédec, S., Gayraud, S., Journal, S., American, N., Society, B., September, N., Biologie, U.R., 2014. Biological traits of stream macroinvertebrate communities: effects of microhabitat, reach, and basin filters. J. North Am. Benthol. Soc. 23, 449–466.
Legendre, P., Legendre, L., 2012. Numerical Ecology-Developments in Environmental Modelling, 3rd ed. Elsevier Science BV, Amsterdam.
Lods-Crozet, B., Castella, E., Cambin, D., Ilg, C., Knispel, S., Mayor-Simeant, H., 2001. Macroinvertebrate community structure in relation to environmental variables in a Swiss glacial stream. Freshw. Biol. 46, 1641–1661.
Lorenz, A.W., Feld, C.K., 2013. Upstream river morphology and riparian land use overrule local restoration effects on ecological status assessment. Hydrobiologia 704, 489–501.
Mazão, G.R., Bispo, P. da C.., 2016. The influence of physical instream spatial variability on Chironomidae (Diptera) assemblages in Neotropical streams. Limnologica 60, 1–5.
McCluskey, A., Lalkhen, A.G., 2007. Statistics II: central tendency and spread of data. Contin. Educ. Anaesthesia. Crit. Care Pain 7, 127–130.
Medhurst, R.B., Wipfli, M.S., Binckley, C., Polivka, K., Hessburg, P.F., Salter, R.B., 2010. Headwater streams and forest management: does ecoregional context influence logging effects on benthic communities? Hydrobiologia 641, 71–83.
Menezes, S., Baird, D.J., Soares, A.M.V.M., 2010. Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. J. Appl. Ecol. 47, 711–719.
Mereta, S.T., Boets, P., Ambelu Bayih, A., Malu, A., Ephrem, Z., Sisay, A., Endale, H., Yitbarek, M., Jemal, A., De Meester, L., Goethals, P.L.M., 2012. Analysis of environmental factors determining the abundance and diversity of macroinvertebrate taxa in natural wetlands of Southwest Ethiopia. Ecol. Inform. 7, 52–61.
Mittermeier, R.A., Myers, N., Mittermeier, C.G., Robles Gil, P., 1999. Hotspots: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions. Agrupación Sierra Madre, S.C., Mexico.
Myers, N., Mittermeier, R., Mittermeier, C., da Fonseca, G., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858.
Naeem, S., 1998. Species redundancy and ecosystem reliability. Conserv. Biol. 12, 39–45.
Ngor, P., Chhuon, K., Prak, L., 2016. Cambodia completes first pilot study of Tonle Sap mollusc fishery. Catch Cult. 22, 4–13.
Nhan, D.K., Phong, L.T., Verdegem, M.J.C., Duong, L.T., Bosma, R.H., Little, D.C., 2007. Integrated freshwater aquaculture, crop and livestock production in the Mekong delta: Vietnam: determinants and the role of the pond. Agric. Syst. 94, 445–458.
Nicola, G.G., Almodóvar, A., Elvira, B., 2010. Effects of environmental factors and predation on benthic communities in headwater streams. Aquat. Sci. 72, 419–429.
Pérez-Quintero, J.C., 2011. Distribution patterns of freshwater molluscs along environmental gradients in the southern Guadiana River basin (SW Iberian Peninsula). Hydrobiologia 678, 65–76.
Palmer, M., Covich, A., Finlay, B., Gibert, J., Hyde, K., Johnson, R., Kairesalo, T., Lake, S., Lovell, C., Naiman, R., Ricci, C., Sabater, F., Strayer, D., 1997. Biodiversity and ecosystem processes in freshwater sediments. Ambio 26, 571–577.
Pan, B., Wang, H., Wang, H., 2014. A floodplain-scale lake classification based on characteristics of macroinvertebrate assemblages and corresponding environmental properties. Limnologica 49, 10–17.
Pearson, R.G., Boyero, L., 2009. Gradients in regional diversity of freshwater taxa. J. North Am. Benthol. Soc. 28, 504–514.
R Core Team, 2013. R: a Language and Environment for Statistical Computing .
Rainboth, W.J., 1996. Species identification field guide for fishery purposes. Fishes of the Cambodian Mekong. FAO, Rome.
Rao, C.R., 1995. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. Qüestiió 19, 23–63.
Roberts, W., 2013. Package labdsv .
Salmah, M.R.C., Al-Shami, S.A., Abu Hassan, A., Madrus, M.R., Nurul Huda, A., 2014. Distribution of detritivores in tropical forest streams of peninsular Malaysia: role of temperature, canopy cover and altitude variability. Int. J. Biometeorol. 58, 679–690.
Sarkkula, J., Kiirikki, M., Koponen, J., Kummu, M., 2003. Ecosystem processes of the tonle sap lake. In: 1st Workshop of Ecotone Phase II. Phnom Penh, Cambodia. .
Sedell, J.R., Rchey, J.E., Swanson, F.J., 1989. The river continuum concept: a basis for the expected ecosystem behavior of very large rivers? Can. Spec. Publ. Fish. Aquat. Sci. 106, 49–55.
Sodhi, N.S., Koh, L.P., Brook, B.W., Ng, P.K.L., 2004. Southeast Asian biodiversity: an impending disaster. Trends Ecol. Evol. 19, 654–660.
Statzner, B., Higler, B., 1985. Questions and comments on the river continuum concept. Can. J. Fish. Aquat. Sci. 42, 1038–1044.
Suhaila, A., Che Salmah, M.R., 2014. Ecology of Ephemeroptera, Plecoptera and Trichoptera (Insecta) in rivers of the Gunung Jerai forest reserve: diversity and distribution of functional feeding groups. Trop. Life Sci. Res. 25, 61–73.
Sweeney, B.W., 1993. Effects of streamside vegetation on macroinvertebrate communities of White Clay Creek in eastern North-America. Proc. Acad. Nat. Sci. Philadelphia 144, 291–340.
Thorp, J.H., Delong, M., 1994. The riverine productivity model: an heuristic view of carbon sources and organic processing in large river ecosystems. Oikos 2, 305–308.
Tonkin, J.D., Death, R.G., Collier, K.J., 2012. Do productivity and disturbance interact to modulate macroinvertebrate diversity in streams? Hydrobiologia 701, 159–172.
Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. The river continuum concept. Can. J. Fish. Aquat. Sci. 37, 130–137.
Vaughn, C., Hakenkamp, C., 2001. The functional role of burrowing bivalves in

freshwater ecosystems. Freshw. Biol. 46, 1431–1446.

Wang, B., Liu, D., Liu, S., Zhang, Y., Lu, D., Wang, L., 2012. Impacts of urbanization on stream habitats and macroinvertebrate communities in the tributaries of Qiangtang River, China. Hydrobiologia 680, 39–51.

Wetzel, R.G., 2001. Limnology: Lake and River Ecosystems, 3rd ed. Academic Press San Diego, CA.

Wilby, A., Lan, L.P., Heong, K.L., Huyen, N.P.D., Quang, N.H., Minh, N.V., Thomas, M.B., 2006. Arthropod diversity and community structure in relation to land use in the Mekong Delta, Vietnam. Ecosystems 9, 538–549.

Zalinge, N.V., Thuok, N., 1998. It's big, unique and important: fisheries in the Lower Mekong Basin: as seen from a Cambodian perspective. Catch Cult. 4, 1–8.

Zalinge, N.V., Degen, P., Pongsri, C., Nuov, S., Jensen, J., Hao, N., Choulamany, X., 2003. The mekong river system. In: Second International Symposium on the Management of Large Rivers for Fisheries Phnom. Phnom Penh, Cambodia. pp. 1–18.

Zilli, F.L., Marchese, M.R., 2011. Patterns in macroinvertebrate assemblages at different spatial scales: implications of hydrological connectivity in a large floodplain river. Hydrobiologia 663, 245–257.

Uniqueness of sampling site contributions to the total variance of macroinvertebrate communities in the Lower Mekong Basin

Ratha Sor[a,b*], Pierre Legendre[c], Sovan Lek[a]

[a] Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS - Université Paul Sabatier –Toulouse 3, 118 route de Narbonne, 31062 Toulouse cédex 4 – France

[b] Department of Biology, Faculty of Science, Royal University of Phnom Penh, Russian Boulevard, 12000, Phnom Penh, Cambodia

[c] Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
[*]Correspondence: Sor, R., e-mail: sorsim.ratha@gmail.com

Abstract

Species co-occurrence and site-specific characteristics have a great influence on biotic community composition at local scales and thus contribute to large variations at broad spatial scales. In this paper, we studied invertebrate communities in 63 river sites of the Lower Mekong Basin (LMB) sampled over 609 thousand km^2. We identified important macroinvertebrate taxa of the component communities (i.e. annelids, crustaceans, mollusks and insects), and key geo-environmental factors that explained the total variance (BD_{Total}) of the communities at large spatial scale. We used the "*Species Contributions to Beta Diversity*" (SCBD) and "*Local Contributions to Beta Diversity*" (LCBD) approaches to partition total beta diversity (BD_{Total}), identified the important macroinvertebrate taxa (those with high SCBD indices), and estimated the uniqueness of sites in community composition (LCBD indices). SCBD indices showed which taxa were the most important in structuring the four component communities: there were 29 insect taxa, which mainly characterized the upstream sites, and 18 mollusk, 7 annelid and 6 crustacean taxa, which all represented the downstream sites. We used linear regression models and variation partitioning to investigate the influence of component communities and of geo-environmental factors on LCBD indices. Our results showed great variation in composition within the LMB (BD_{Total} = 0.80 on a 0-to-1 scale). Five sites of the main channel exhibited significant uniqueness (LCBD indices) in community composition. One of them was a hotspot location occupied by a community with exceptional taxonomic composition, which should be protected. Four other sites were degraded by human activity and in need of restoration. Variation partitioning indicated that LCBD indices are well explained by two main component communities (mollusks and insects, adjusted R^2 = 0.84), and by water conductivity, river depth and Secchi depth (adjusted R^2 = 0.26). The two sets of explanatory factors jointly explained a fairly large fraction of the LCBD variation (adjusted R^2 = 0.24**)**. LCBD variation responded more to the composition of component communities and environmental factors than to geographical factors. The uniqueness in community composition of the sites that we estimated provides useful ecological information, which could support restoration and conservation planning for the LMB.

Keywords: Beta diversity, local contribution to beta diversity, species contribution to beta diversity, annelids, crustaceans, mollusks, insects, environmental degradation, river restoration and management

1. Introduction

The variation in community composition among sites, or beta (β) diversity (Legendre and De Cáceres, 2013; Whittaker, 1960), is of primary interest to community ecology. Beta diversity is an important component of biodiversity as it links local (α) to regional (γ) diversity, and it varies as a function of the spatial scales and gradients of the study areas (Anderson et al., 2011; Legendre and Legendre, 2012; Whittaker, 1972, 1960). Therefore, understanding the variation in species composition among sites, i.e. β diversity, enables community ecologists to disclose evolutionary and ecological processes at work in a community of interest (Valdujo et al., 2013), by analyzing and testing such processes in a way that indicates how they affect and maintain biodiversity in the ecosystem (Legendre and De Cáceres, 2013).

Biotic interactions in communities, e.g. intra-specific competition, have been reported to influence β diversity (Matthiessen and Hillebrand, 2006; Valdujo et al., 2013). In identical ecological patches, interactions among species within their own taxonomic group or between different taxonomic groups can lead to different patterns of β diversity (Hillebrand and Blenckner, 2002; Tonkin et al., 2015), and thus affect ecosystem functioning. Environmental gradients, habitat heterogeneity (López-González et al., 2015), and natural and human-derived disturbances (Lamy et al., 2015; Legendre and Salvat, 2015) have been shown to also influence β diversity. For aquatic macroinvertebrates, β diversity is mainly related to drainage basins and within-stream environmental factors, while it has been reported not to be significantly related to habitat degradation, eutrophication, longitude and altitude (Friberg et al., 2010; Md Rawi et al., 2013). However, longitude and altitude have been found to be substitute variables (proxies) for major drivers patterning β diversity of macroinvertebrates at broad geographical scales (J. Wang et al., 2012). The environmental variables related to the geographical proxies may play important roles in structuring the broad-scale pattern of β diversity in a given region.

Several papers have reported patterns of β diversity in tropical ecosystems for plants and vertebrate animals (e.g. Legendre et al., 2009; López-González et al., 2015; Mena and Vázquez-Domínguez, 2005; Wearn et al., 2016). β diversity of macroinvertebrates has also recently been analyzed by several authors (e.g. Costa and Melo, 2008; Leigh and Sheldon, 2009; Ligeiro et al., 2010), but only a few studies have taken place in South-East Asia (e.g. Al-Shami et al., 2013, Salmah et al., 2014). As the ecosystems in that region are highly endangered and heavily impacted by human disturbances (Salmah et al., 2014; Sodhi et al., 2004; Strayer and Dudgeon, 2010), assessing the patterns of macroinvertebrate β diversity and their relationships to geo-environmental factors and to related biotic communities is urgently needed.

The Mekong River Basin is divided into Upper and Lower Mekong Basins (LMB). The LMB, covering an area of about 609,000 km^2 (77% of the whole basin) (Zalinge et al., 2003), includes portions of four densely populated countries: Thailand, Laos, Cambodia and Vietnam. This basin harbors diversified communities of fish and invertebrates, forming biologically important food webs that support high biodiversity (Sodhi et al., 2004). Many aquatic taxonomic groups such as fishes, mollusks, crustaceans and insects are highly dependent on this basin as a breeding ground (Davidson et al., 2006; Zalinge and Thuok, 1998). In spite of high suspected biodiversity in the LMB, the β diversity and community patterns of its aquatic taxonomic groups, particularly the macroinvertebrates, have seldom been studied. The biomonitoring surveys conducted by the Mekong River Commission (MRC) represent the only major work conducted on aquatic macroinvertebrates in the LMB. In this study, we used this bio-monitoring data to explore the β diversity pattern of aquatic macroinvertebrates. Analyzes of this dataset, collected from sites sampled over 5 successive years (2004-2008), should significantly contribute to increase our scientific knowledge of the LMB.

Beta diversity can be computed in different ways (Koleff et al., 2003; Whittaker, 1960). A classical approach is to compute β diversity as $\beta = \gamma/\bar{\alpha}$, where γ is the total number of species in a given region and $\bar{\alpha}$ is the average number of taxa for a sample set within the region (Whittaker, 1960). This classical measurement is still preferred by many authors (Higgins, 2010; Jost, 2007; Sor et al., 2015) although new approaches have been developed (Anderson et al., 2011; Legendre et al., 2005; Legendre and De Cáceres, 2013).

In this study, we used the total variance of the macroinvertebrate communities among the study sites of the LMB as a measure of beta diversity (BD_{Total}) and partitioned it into "*Local Contributions to Beta Diversity*" (LCBD) and "*Species Contributions to Beta Diversity*" (SCBD) (Legendre and De Cáceres, 2013). We identified the important taxa contributing most to total β diversity, i.e. those with high among-site variance, as well as the geo-environmental factors that were associated with the macroinvertebrate communities throughout the sites. In addition, we investigated the influence of the LCBD indices of the component communities (i.e. annelids, crustaceans, mollusks and insects) on the LCBD indices of the global macroinvertebrate community composition (including all component communities). Our questions of interest are the following: 1) Is there a moderate or a large amount of variation in macroinvertebrate community composition among the sites in the LMB? 2) What are the taxa that contribute most to the total β diversity? We expect the important taxa of annelids, crustaceans and mollusks, measured as richness and abundance, to be associated with sites located downstream, whereas the important taxa of insects should be associated with sites located farther upstream, as has been shown by Arscott et al. (2005) and Królak and Korycińska (2008). 3) Are there sites that have exceptionally unique taxonomic compositions? We hypothesize that some sampling locations exhibit significant uniqueness in taxonomic composition. 4) What are the geo-environmental conditions that characterize the sites with significant LCBD indices? We expect the LCBD indices to increase with river width and pH, following the β diversity patterns found in tropical streams in Malaysia (Al-Shami et al., 2013), and decrease with latitude and altitude, following the β diversity patterns observed in major geographical diversity gradients (J. Wang et al., 2012). 5) What are the component communities that mainly influence the LCBD indices of the global macroinvertebrate communities?

2. Materials and methods

2.1 Macroinvertebrate and geo-environmental variables collection

From 2004 to 2008, the Mekong River Commission (MRC) conducted biomonitoring surveys and sampled macroinvertebrates at 60 sites along the LMB once a year in March during the dry season (Fig. 1). To harmonize the data being collected, the sampling locations were selected from different habitats such as those in or close to villages or towns, at rivers with substantial shipping, next to crop fields and meadows with livestock, upstream or downstream of dams or weirs, and at more pristine areas surrounded by forest with only few houses. At each sampling site, benthic macroinvertebrates and geo-environmental variables were collected at the same time. For the detailed information on the collection process, we refer to Sor et al. (2017).

In 2008, 3 sampling sites were sampled farther away from their original sampling coordinates, and thus they were regarded as new sampling sites (see Appendix T1 in Part I: Synthesis). Therefore, we considered a total of 63 sampling sites in the present study.

2.2 Data processing and statistical methods

For the 63 sampling sites, 108 samples of biological and geo-environmental variables were available. Due to unequal sampling efforts, a small number of sites were sampled only once, twice or thrice during the 5-year sampling period. Since this is the first survey of macroinvertebrates ever conducted in the LMB and the sampling protocol insured that the collected samples were comparable among sites, these data are important to obtain a first assessment of beta diversity. Therefore, we used median values from data collected on macroinvertebrate and geo-environmental variables to represent each site in our analyzes, as suggested for small sample size by McCluskey and Lalkhen (2007). The community composition data was partitioned into a global macroinvertebrate community data table (including all component communities), and component community data tables (for annelid, crustacean, mollusk and insect communities).

The community composition data were Hellinger-transformed at the beginning of the analyzes (Legendre and Gallagher, 2001; Legendre and Legendre, 2012). For Hellinger-transformed data, the total variance, or total β diversity (BD_{Total}), of a community composition data table is an index between 0 and 1, and it can be partitioned into local contribution (LCBD) and species contribution (SCBD) indices. An LCBD value is an index showing the degree of uniqueness in taxonomic composition in each site, computed as the relative contribution of a site to BD_{Total}, so that the LCBD indices sum to 1, whereas an SCBD index shows the relative degree of variation of a taxon across all sites. The BD_{Total}, LCBD and SCBD indices were computed using the function "beta.div" available in the *adespatial* package in R (Dray et al., 2016). The Hellinger transformation was used because the corresponding Hellinger distance is one of the dissimilarity functions admissible for beta diversity analyzes (Legendre and De Cáceres, 2013; Legendre and Gallagher, 2001); it does not give high weights to the rare species. To identify significant uniqueness in taxonomic composition of the sampling sites, the LCBD indices were tested for significance against a significance level $\alpha = 0.05$. The p-values were corrected for multiple testing using the Holm correction to reduce the experimentwise type I error rate of multiple tests. In addition to LCBD, Hellinger-transformed data also allow researchers to compute SCBD indices; this is not allowed by most other admissible dissimilarity functions (Legendre and De Cáceres, 2013; Legendre and Gallagher, 2001). In the following paragraphs, BD_{Total}, LCBD and SCBD designate the indices of the global macroinvertebrate communities, whereas BD_{ATotal}, BD_{CTotal}, BD_{MTotal}, BD_{ITotal}, and $LCBD_{A}$, $LCBD_{C}$, $LCBD_{M}$ and $LCBD_{I}$ designate the BD_{Total} and LCBD indices for annelid, crustacean, mollusk and insect communities, respectively.

SCBD indices that were higher than the mean of SCBD values identified the taxa that were the most important contributors to BD_{Total}. Before associating these important taxa with the geo-environmental factors, we normalized these variables using the indications provided by function "boxcoxfit" in the *geoR* package in R (Ribeiro Jr and Diggle, 2015). The Box-Cox transformation was performed on the geo-environmental variables because this transformation attempts to normalize the variables, thus meeting the assumptions of linear models and residuals' normal distributions (Ahola et al., 2011). Then, we conducted a Redundancy Analysis (RDA, Legendre and Legendre 2012) on the Hellinger-transformed abundance data. To identify which component community was more related to which part (downstream or upstream) of the LMB, we computed Pearson correlations between the richness (number of taxa) and abundance (number of individuals) of the important taxa pertaining to each component community, on the one hand, and to the geographical factors on the other hand.

We independently ran simple and multiple regression analyzes to determine which, among the geographical and environmental variables, mainly accounted for the variation of the LCBD indices. To identify the strength of the regression models, we computed stepwise selection with the Akaike Information Criterion (AIC). The models having the lowest AIC and highest adjusted R^2 were considered to have the strongest influence on the LCBD indices. To investigate the influence of component communities on the LCBD indices, we computed $LCBD_{A}$, $LCBD_{C}$, $LCBD_{M}$ and $LCBD_{I}$, and regressed the LCBD indices (for the global communities) on the LCBD indices of the four component communities. We computed four types of linear regression models: 1) Simple regression models, e.g. $LCBD \sim LCBD_{A}$; 2) 2-component multiple regression, e.g. $LCBD \sim LCBD_{A} + LCBD_{C}$; 3) 3-component multiple regression, e.g. $LCBD \sim LCBD_{A} + LCBD_{C} + LCBD_{M}$ and 4) all-component multiple regression, $LCBD \sim LCBD_{A} + LCBD_{C} + LCBD_{M} + LCBD_{I}$. Model selection, based on the AIC, was conducted to obtain a descriptive assessment of the components that contribute most to the variation of the global LCBD indices. Finally, variation partitioning was applied to quantify the variance that main component communities and the significant geo-environmental best explain the variation of the global LCBD indices. All statistical analyses were performed in R (R Core Team, 2013).

3. Results

3.1 General macroinvertebrate composition and environmental variables

In total, 21,810 individuals representing 299 taxa and 90 families were identified in the dataset (see Appendix T2 in Part I: Synthesis). Taxonomic richness was highest at the Mekong delta sites (Fig. 1). Among the taxa, 32 belonged to annelids (2,672 individuals), 38 to crustaceans (2,054), 98 to mollusks (10,603) and 131 to insects (6,481). The most common families of annelids were Naididae (47% of occurrence) and Nereididae (16%); of crustaceans were Palaemonidae (26%) and Corophiidae (16%) and of mollusks were Unionidae (18%), Corbiculidae (14%), Viviparidae (12%) and Stenothyridae (9%). Insect communities were characterized by Diptera (28%), Ephemeroptera (24%), Odonata (17%) and Trichoptera (15%).

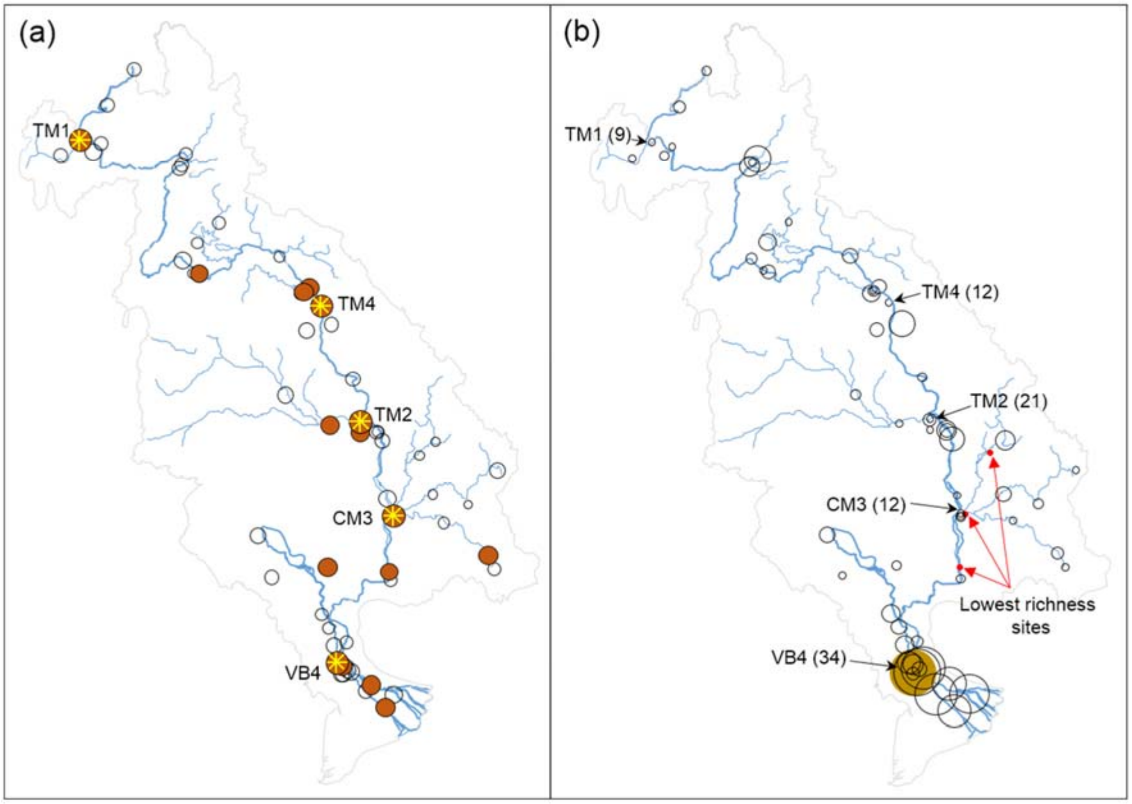


Fig. 1. Map of the sampling sites in the LMB. (a) LCBD indices with significant p-values uncorrected (brown dots) and corrected for multiple testing by applying Holm correction (shaded dots with star); open circles: non-significant LCBD indices. (b) Richness (number of taxa) of the sampling sites. Three red dots indicated with red arrows: lowest richness, 6; large shaded circle: highest richness, 74. The richness for the five sites having significant LCBD indices is shown in the parentheses. The sizes of the circles are proportional to LCBD (a) or richness (b) values.

Three taxa were most widely distributed; two belonged to insects: *Ablabesmyia* sp. (73% occurrence) and *Polypedilum* sp. (70%) and one was a mollusk, *Corbicula tenuis* (67%). In addition to being widely distributed, these 3 taxa were among the top 10 most abundant. Of the total individuals, *Ablabesmyia* sp. accounted 2.9%, *Polypedilum* sp. for 3.8%, whereas the 3 most abundant species, *Corbicula leviuscula*, *Limnoperna siamensis* and *Corbicula tenuis*, accounted for 8.4%, 6.1% and 5.8%, respectively. The data on taxonomic richness and abundance, and the environmental variables are summarized in Table 1.

Table 1 Observed environmental factors, richness (number of taxa) and abundance (number of individuals) of macroinvertebrates across the 63 sampling sites.

Variables	Unit	Min	Max	Mean	Standard deviation
Altitude	m	3	546	127	132
Water temperature	° C	17	31	27	3
Dissolved oxygen	mg/L	2.7	9.3	7.4	2.6
Water conductivity	mS/m	3.9	66.6	17.9	11.9
River width	m	11	1629	467	466
River depth	m	0.4	15.0	4.8	3.9
Secchi depth	m	0.2	3.0	0.9	0.6
pH	—	6.8	8.4	7.6	0.6
Richness	taxa/sample	6	74	23	16
Abundance	individuals/sample	13	1997	315	396

3.2 Beta diversity and important taxa - habitat relationship

The total β diversity of macroinvertebrates in the LMB was BD_{Total} = 0.80, and there was a total of 60 taxa contributing most to the BD_{Total}. The value of BD_{Total} is very high, considering that the maximum that can be obtained for Hellinger-transformed data is 1, when all sites have entirely different species compositions. This great variation was also observed for each component community: annelids (BD_{ATotal} = 0.72), mollusks BD_{MTotal} = 0.78) and insects (BD_{ITotal} = 0.74), excepted for crustaceans (BD_{CTotal} = 0.38). Over all 299 taxa, 60 important taxa had SCBD indices larger than the mean SCBD (0.003), 29 of which belong to insects, 18 to mollusks, 7 to annelids and 6 to crustaceans (see Table 3.3 in Part I: Synthesis). The SCBD values are small because the SCBD indices are relative to the total sum of squares in the community composition table and sum to 1. SCBD indices indicate taxa that have the highest variance across sites. The 3 highest SCBD indices belonged to insect taxa: *Polypedilum* sp. (0.054), *Ablabesmyia* sp. (0.039), *Cryptochironomus* sp. (0.037), followed by *Corbicula tenuis* (mollusk, 0.037), *Goeldichironomus* sp. (insect, 0.035) and *Corbicula leviuscula* (mollusk, 0.034).

Based on the correlation analyzes, the richness and abundance of the important taxa of annelids, crustaceans and mollusks were significantly and negatively correlated with latitude and altitude; the richness and abundance of the important taxa of insects were significantly and positively correlated with latitude and altitude (Table 2). As the Mekong River generally runs from north to south, decreasing latitude and decreasing altitude are both associated with going downstream.

Table 2 Pearson correlation coefficients between the normalized geographical factors and richness and abundance of each component community among the important taxa. R: richness (number of taxa), A: abundance (number of individuals), LONG: longitude (m), LAT: latitude (m), ALT: altitude (m). Significant relationships are marked with stars.* p < 0.05, ** p < 0.01, *** p < 0.001

	Annelids		Crustaceans		Mollusks		Insects	
	R	A	R	A	R	A	R	A
LONG	0.15	0.06	0.03	0	-0.01	-0.04	-0.03	-0.05
LAT	-0.58**	-0.36**	-0.70***	-0.25*	-0.65***	-0.35**	0.47***	0.33**
ALT	-0.65***	-0.42***	-0.65***	-0.26*	-0.69***	-0.35**	0.41***	0.25*

The first two axes of the RDA model (Fig. 2a, b) accounted for 15.7% of the total variance of the community of important taxa. Axis 1 (9.4%) of the plot showed the river gradients opposing river depth and river width (left of the plot), which were strongly associated with communities in the Mekong delta, to altitude and latitude, which were associated with communities in the upstream sites. Along axis 2 (6.3%), the communities from some tributaries and main channel sites located upstream in the LMB (Thailand and Laos) were found in the positive part of the axis and associated with high values of dissolved oxygen, while many other tributary sites were found in the lower part of the axis and related to high values of Secchi depth.

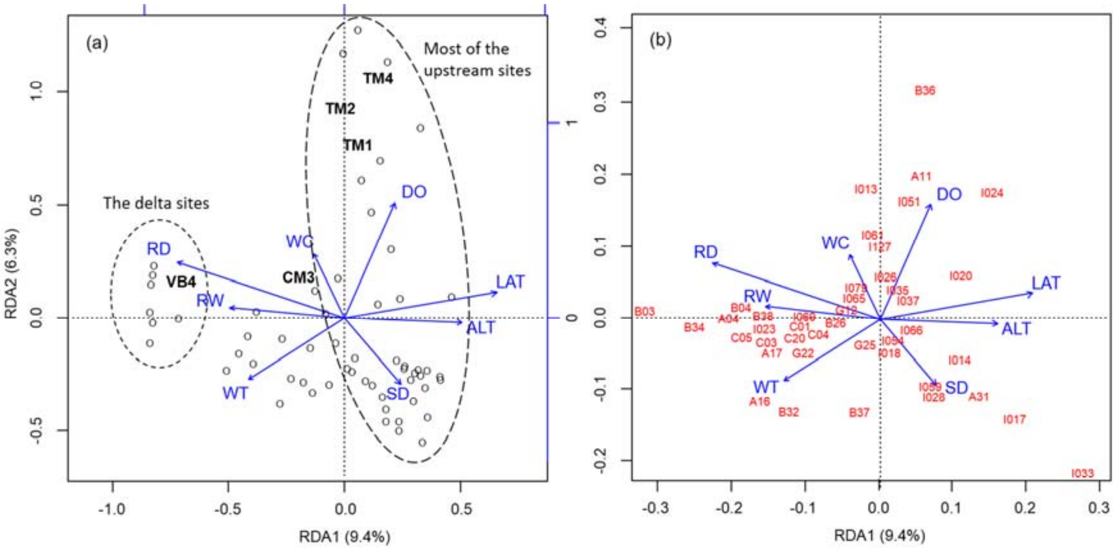


Fig. 2. RDA ordination plots showing the association of taxa having SCBD higher than the mean value (i.e. important taxa) with the geo-environmental factors. (a) Plot showing the sites and geo-environmental factors. VB4, CM3, TM1, TM2 and TM3 in bold are the sites with significant LCBD indices. LAT: latitude, ALT: altitude, WT: water temperature, DO: dissolved oxygen, WC: water conductivity, RW: river width, RD: river depth, SD: Secchi depth. (b) Plot showing taxa with high SCBD indices and geo-environmental factors for the same analyzes. The first letter of each taxon name represents the component of macroinvertebrate communities (A: annelid, C: crustacean, B and G: mollusk, and I: insect). Geo-environmental variables and taxa that have small loading score were removed to improve legibility.

3.3 Uniqueness in taxonomic composition and its association with geo-environmental factors

Five sampling sites exhibited significant global LCBD indices at the p = 0.05 level after Holm correction for multiple testing. These sites were CM3 (LCBD = 0.023, p = 0.006), TM2 and TM4 (LCBD = 0.021, p = 0.006), and TM1 and VB4 (LCBD = 0.020, p = 0.012)

(Fig. 1). CM3 also exhibited significant LCBD indices for all component communities ($LCBD_A$ = 0.026, p = 0.032; $LCBD_C$ = 0.038, p = 0.006; $LCBD_M$ = 0.023, p = 0.006; $LCBD_I$ = 0.027, p = 0.006). LCBD indices of the global communities and of component communities, which indicate the uniqueness in taxonomic composition at the sites, are provided in Appendix T3 in Part I: Synthesis. LCBD values are scaled to add up to 1 over the whole study; the mean LCBD value in this study was thus 1/63 = 0.016.

Results of the simple and multiple regressions between the global LCBD indices and geo-environmental variables are shown in Table 3. Only water conductivity, river depth and Secchi depth were significantly associated with the global LCBD indices, and these variables remained significant after the stepwise selection of the model (Table 3). Water conductivity and river depth were positively associated and accounted for 14% and 12%, and Secchi depth was negatively associated and accounted for 6% of the variation of LCBD indices (Table 3).

Table 3 Contribution of geo-environmental factors to the variation of LCBD indices of global macroinvertebrate communities. The model showing the strongest contribution is in bold. WT: water temperature, DO: dissolved oxygen, WC: water conductivity, RW: river width, RD: river depth, SD: Secchi depth, LONG: longitude (m), LAT: latitude (m), ALT: altitude (m).

Model	Adjusted R^2	AIC	Stepwise selection	F-stat	p of the model
Environmental factors					
LCBD ~ WT	--	-551.2	--	0.01	0.956
LCBD ~ DO	--	-551.5	--	0.3	0.609
LCBD ~ WC	0.135	-561.4	--	10.7	0.002
LCBD ~ RW	0.010	-552.9	--	1.6	0.206
LCBD ~ RD	0.120	-560.2	--	9.5	0.003
LCBD ~ SD	0.061	-556.2	--	5.1	0.029
LCBD ~ pH	--	-551.5	--	0.3	0.590
LCBD ~ all factors	0.235	-563.6	WC+RD+SD	3.7	0.002
LCBD ~ WC+RD+SD	**0.260**	**-569.4**	**WC+RD+SD**	**8.3**	**<0.001**
Geographical factors					
LCBD ~ LONG	--	-551.8	--	0.6	0.449
LCBD ~ LAT	0.001	-552.3	--	1.0	0.314
LCBD ~ ALT	0.010	-552.9	--	1.6	0.207
LCBD ~ all factors	--	-549.7	--	0.8	0.511

3.4 Influence of component communities and environmental conditions on global LCBD indices

Among the simple regression models, $LCBD_I$ indices were the strongest determinants of the variation of the global LCBD indices. In multiple regressions (2-, 3-, and 4-components), the combination of $LCBD_I$ and $LCBD_M$ indices best explained the variation of the global LCBD indices (Table 4). The detailed results of the regression models are shown in Table 4. Based on the variation partitioning results (Fig. 3), the variation of the global LCBD indices was well explained by that of the $LCBD_I$ and $LCBD_M$ (adjusted R^2 = 0.84), whereas the three significant environmental variables explained a smaller fraction (adjusted R^2 = 0.26) of the variation of the global LCBD indices. For the most part, the environmental variable fraction represented explanation shared with $LCBD_I$ and $LCBD_M$ (adjusted R^2 = 0.24).

Table 4 Contribution of the composition of component communities to the LCBD indices of the global macroinvertebrate communities. The model showing the strongest contribution is in bold. $LCBD_A$, $LCBD_C$, $LCBD_M$, $LCBD_I$ are the LCBD indices of each component community (annelids, crustaceans, mollusks and insects, respectively).

Model	Adjusted R^2	AIC
LCBD ~ $LCBD_A$	0.331	-577.6
LCBD ~ $LCBD_C$	0.053	-555.7
LCBD ~ $LCBD_M$	0.241	-569.7
LCBD ~ $LCBD_I$	0.776	-646.5
LCBD ~ $LCBD_A$+$LCBD_C$	0.358	-579.3
LCBD ~ $LCBD_A$+$LCBD_M$	0.429	-586.6
LCBD ~ $LCBD_A$+$LCBD_I$	0.778	-646.3
LCBD ~ $LCBD_C$+$LCBD_M$	0.264	-570.7
LCBD ~ $LCBD_C$+$LCBD_I$	0.772	-644.5
LCBD ~ $LCBD_M$+$LCBD_I$	**0.843**	**-668.0**
LCBD ~ $LCBD_A$+$LCBD_C$+$LCBD_M$	0.444	-587.4
LCBD ~ $LCBD_A$+$LCBD_M$+$LCBD_I$	0.841	-666.2
LCBD ~ $LCBD_C$+$LCBD_M$+$LCBD_I$	0.841	-666.1
LCBD ~ $LCBD_A$+$LCBD_C$+ $LCBD_M$+ $LCBD_I$	0.834	-664.3

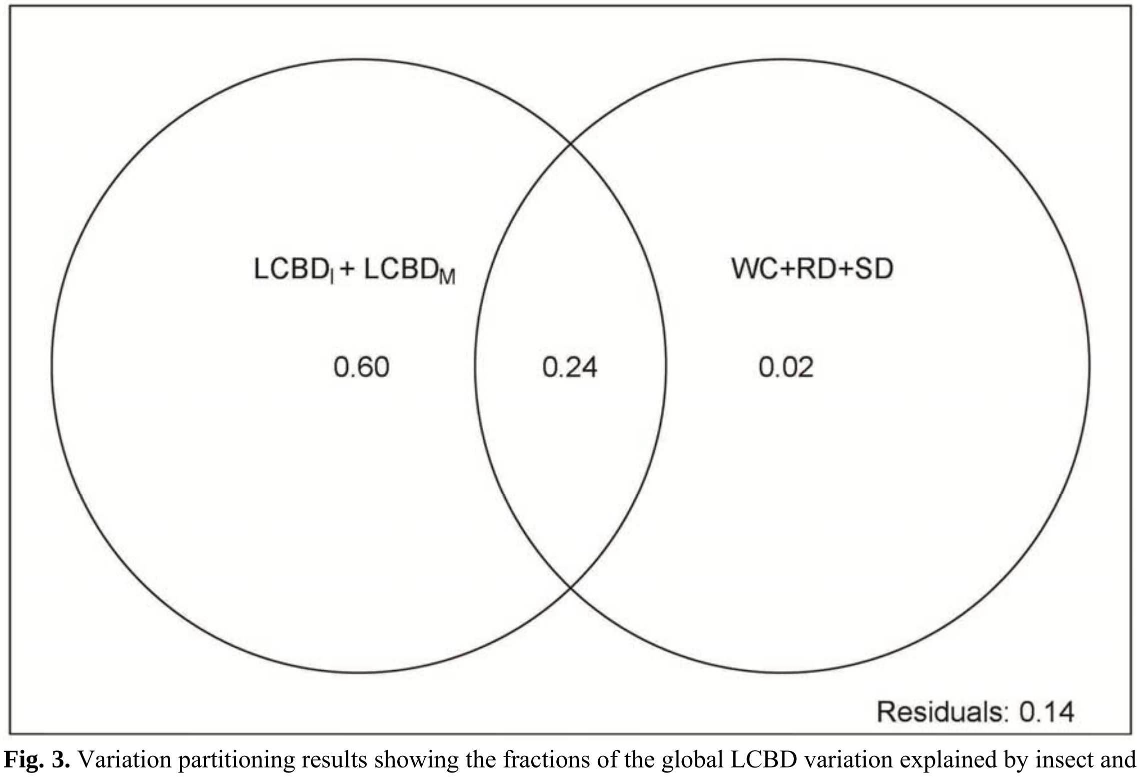


Fig. 3. Variation partitioning results showing the fractions of the global LCBD variation explained by insect and mollusk communities ($LCBD_I$ + $LCBD_M$), by the three environmental variables (WC: water conductivity, RD: river depth, SD: Secchi depth), and by the two sets of explanatory factors.

4. Discussion

4.1 Macroinvertebrate communities of the LMB

Macroinvertebrate communities in the LMB has been scarcely studied, in particular at a broad large spatial scale. To our knowledge, most of the existing studies were conducted at local sub-watershed scales of the LMB (e.g. Clavier et al., 2015; Cuong et al., 2016; Getwongsa et al., 2010; Quang et al., 2013), except for a macroinvertebrate pilot study by Pathoumthong and Vongsombath (2007), which was conducted over thirteen sampling sites of the LMB and reported 218 macroinvertebrate taxa. The number of macroinvertebrate species (299) identified in the present study is the largest ever reported from the basin. Each component (i.e. annelids, crustaceans, mollusks and insects) comprised a higher number of taxa than previous reports of field studies (Table 5). Moreover, we found that 36 macroinvertebrate families were represented by only one species (see Appendix T2 in Part I: Synthesis). This indicates that the LMB could support numerous rare or endemic taxa as found for mollusks that at least 111 species are endemic to the LMB (Köhler et al., 2012).

Table 5 The number of recorded taxa of different taxonomic groups of macroinvertebrates recorded in the present study and in previous reports. The number of arthropod taxa in the parenthesis is the combination of crustacean and insect taxa.

Studied area	Number of reported taxa						References
	Annelids	Crustaceans	Mollusks	Insects	Arthropods	Overall	
The LMB	32	38	98	131	(169)	299	Present study
	—	—	—	—	—	218	Pathoumthong and Vongsombath (2007)
Thailand	—	—	—	—	—	164	Getwongsa et al. (2010)
Laos	3	5	13	86	—	109	Clavier et al. (2015)
Cambodia	—	—	22	—	—	—	Ngor et al. (2016)
	—	—	—	17	—	—	Pauly (2016)
Vietnam (delta)	16	26	56	27	—	125	Quang et al. (2013)
	—	—	—	—	578	—	Cuong et al. (2016)
Indo-Burma	—	—	~146	—	—	—	Köhler et al. (2012)

4.2 Variation of important taxa and their relationship to habitat characteristics

As stated in our second research question, the annelids, crustaceans and mollusks that have high SCBD indices are more abundant in the downstream part of the LMB, as reflected by the strong negative correlation with latitude and altitude (Table 2). Unsurprisingly, this result also supports previous studies (Arscott et al., 2005; Collier and Lill, 2008; B. Wang et al., 2012). The important taxa included *Corbicula leviuscula* (code B34 in Fig. 2b), *C. lamarckiana* (B32), *Limnoperna siamensis* (B03) and *Branchiura sowerbyi* (A16), which mostly and abundantly occurred in the downstream part. However, an insect (i.e. *Cricotopus* sp., I023) was also an important taxon found in the delta. This could be due to the fact that, many species in this genus are capable of withstanding low oxygen concentrations, are resistant to heavy metals, able to withstand high salt concentrations or pollution and can feed on rice (Boesel, 1983; Sinclair and Gresens, 2008), which are all characteristics observed in the delta.

Most of the important taxa characterizing the upstream sites belong to insects. In tropical as well as temperate regions, clear water and high values of dissolved oxygen are mostly found in tributaries and upstream sites, which are mainly preferred by insect taxa (Collier and Lill, 2008; Dobson et al., 2002; Królak and Korycińska, 2008). Of the 3 taxa found with the highest SCBD indices (*Polypedilum* sp.; *Ablabesmyia* sp. and *Cryptochironomus* sp.), *Polypedilum* sp. (code I033 in Fig. 2b) and *Ablabesmyia* sp. (I017) were highly associated with sites having high values of SD (Fig. 2b), which were mostly observed in tributaries. *Cryptochironomus* sp. (I024) and other taxa such as *Bezzia* sp. (I013) and *Anagenesia* sp. (I051) were more associated with high values of dissolved oxygen, which occurred at three of the sites with significantly unique taxonomic composition (TM1, TM2, TM4). Surprisingly, *Corbicula* sp. (B36) and Oligochaeta sp. (A11) were also more associated with these sites. These two taxa may have important taxonomic and ecological value because they were restricted to the main channel shared by Thailand and Laos and its nearby sites.

4.3 Beta diversity and uniqueness in community composition

The β diversity of macroinvertebrates in tropical river systems, particularly in South-East Asia, has not been extensively studied (Boyero et al., 2009; Dudgeon, 2008). Furthermore, the published papers (Al-Shami et al., 2013; Salmah et al., 2014) did not estimate the β diversity of macroinvertebrates as the total variance (BD_{Total}) of the communities found at the sampling sites and computed the contributions of individual sampling sites (LCBD indices) to total β diversity. This measure (i.e. BD_{Total}) quantified "the variation in macroinvertebrate composition among studied sites in the LMB", to which is referred as β diversity by ecologists (Anderson et al., 2011; Legendre et al., 2005; Legendre and De Cáceres, 2013; Whittaker, 1972, 1960). The BD_{Total} computed here is an independent derived quantity that can certainly measure community differentiation of studied taxa, and thus more suitable to analyse beta diversity of macroinvertebrates in the LMB, when compared to the classical approach (i.e. the additive or multiplicative) which is dependent on alpha and gamma diversity. The great variation of macroinvertebrate composition ($BD_{Total} = 0.80$) found may reveal complex evolutionary and ecological processes operating at a site-to-global spatial scale of the LMB.

The contributions of sampling sites (LCBDs) to BD_{Total} can indicate the ecological uniqueness of each sampling site in terms of community composition and provide valuable information on the level of habitat degradation of sampling sites. These ecological indications can be used to support ecological assessments, restoration and conservation planning of the LMB. For example, we found that sites with large LCBD indices, which are the most different from the centroid of the distribution of the sites in a PCA ordination and hence the most interesting to examine in detail, mostly occurred along the main channel of the LMB. In particular, the 5 sites that had significant uniqueness in species composition (after Holm correction) occurred along the main channel and not in tributaries (Fig. 1a). The discussion on what triggered these sites to have higher degrees of uniqueness in species composition than others is provided in the following paragraphs.

4.4 Environmental factors responsible for uniqueness in community composition

Three environmental factors were found to be positively (water conductivity and river depth) or negatively (Secchi depth) associated with the degree of site uniqueness in taxonomic composition (Table 3). These 3 factors collectively explained 26% (adjusted $R^2 = 0.26$) of the variance in degrees of site uniqueness in taxonomic composition (global LCBD indices). However, our results did not find the types of relationships between β diversity and geo-environmental factors found in previous studies conducted over smaller areas (Al-Shami et al., 2013; J. Wang et al., 2012). This could be due to the dominant effect of anthropic pressure, which is spread along the LMB (Dao et al., 2010; Kudthalang and Thanee, 2010).

Previous papers have shown that conductivity had a positive influence on macroinvertebrate diversity (Lods-Crozet et al., 2001; Rizo-Patrón V. et al., 2013). However, we found that most of the sites with significant uniqueness in taxonomic composition had low taxonomic richness. High values of conductivity were mostly measured in the main channel sites (e.g.

sites TM2, TM4 and the nearby sites) where they receive runoffs and discharge of urban wastewaters from intensified agriculture from surrounding river basins and cities (Dao et al., 2010; Kudthalang and Thanee, 2010; Sor et al., 2017), and consequently lead to high conductivity (Wetzel, 2001). When sources of pollution (i.e. high concentration of inorganic dissolved solids) enter the rivers, only pollution- or disturbance-tolerant taxa (e.g. Oligochaeta, Chironomidae (Diptera) and Gastropoda) can resist (Feld and Hering, 2007; B. Wang et al., 2012). Pollution is the connection between high conductivity and low taxonomic richness.

Deep rivers (i.e. with high values of river depth) that have low Secchi depth can be considered proxies for anthropic activities (Baird and Flaherty, 2005; Dao et al., 2010), which is why they appear to influence LCBD indices. For example, we found small values of LCBD indices at most of the tributary sites where clear water and low pollution are observed, whereas large values of LCBD indices (e.g. > mean LCBD value) were found at sites with high river depth (e.g. most sites in the delta) and at other sites along the main channel of the upper part of the basin (Fig. 2a, see Appendix T3 in Part I: Synthesis) where high levels of anthropic disturbance were observed. A clear evidence of the association between anthropic activities and high LCBD indices is found at the sites with significant LCBD indices (e.g. CM3, TM1, TM2 and TM4), all of which receive a moderate to high pressure of human impacts. Site CM3 is surrounded by houses, animal wastes and rubbish disposal. Site TM1 seems to be in a very high pressure area since it is opened to many anthropic activities such as animal and human waste disposal, artificial bank creation, local markets, dense population (~10,000 inhabitants), constructions, fishing and boat traffic. Sites TM2 and TM4 are also exposed to waste disposal and fishing, floating houses, tourism (TM2) and agriculture (TM4) (Dao et al., 2010; Kudthalang and Thanee, 2010; Sor et al., 2017). As a result, these sites supported low numbers of taxa (TM1, 9 taxa; TM2, 12; TM4, 21; CM3, 12), which indicate that ecological restoration is needed for these sites and their surroundings.

On the other hand, site VB4 has a significant LCBD value with a moderate number of taxa (34 taxa); the site with highest richness in our study was VB1 (74 species), located close to VB4. VB4 is located at the border between Cambodia and Vietnam and comprises a set of natural land cover (e.g. wood-, shrub-, grass-, inundated and wetland) on the west side of the river. Although the other side of the river has some anthropic activities including houses, fishing and small-scale business, VB4 still had a unique and rich taxonomic composition (Fig. 1a, b). Thus, VB4 and the surrounding sites/areas, particularly on the west side with natural land covers, may have high conservation value. Sites with high LCBD values may have high or low species richness, as shown in Legendre and De Cáceres (2013) and found in the research reported here.

4.5 Influence of component communities and environmental conditions on global LCBD indices

The variation partitioning indicated the most striking relationship (adjusted R^2 = 0.84) between the uniqueness in taxonomic composition of the macroinvertebrate communities (global LCBD indices) and the combination of uniqueness in taxonomic compositions of the mollusk and insect communities ($LCBD_M$ + $LCBD_I$ indices). Note that the global LCBD indices are not simply the sum of the component community LCBD indices; LCBD indices are computed separately for the global study and each component group as the squared distances of the sites to the multivariate ordination centroid. However, the degree of uniqueness in taxonomic composition of macroinvertebrate communities (global LCBD indices) is expected to be contributed by the component communities. In the LMB, LCBDs of mollusk and insect communities, which had a higher total variation (BD_{MTotal} = 0.78 and

BD_{ITotal} = 0.74, respectively), explained most of the global LCBD variation because these two groups had higher abundances and wider distributions than the annelid and crustacean communities, which had lower total variation (BD_{ATotal} = 0.72 and BD_{CTotal} = 0.38, respectively). De'ath (2002) and Davidson et al. (2010) mentioned that taxa with low richness and low occurrence explained less variance of the community composition, and this is similar to our findings for the annelid and crustacean communities.

Co-occurrence among different component communities can directly or indirectly constrain the spatial distributions and the taxonomic abundance of the component communities (Miller, 1994; Wootton, 1994). Predators (e.g. Odonate taxa and some of the Diptera) may prey upon the taxa of other taxonomic communities, and thus affect the taxonomic occurrence and abundance of the global macroinvertebrate communities. Golfieri et al. (2016) reported that the abundance of the Odonates is closely linked to the abundance of their prey in the ecosystems. However, the Odonates preferring high water quality in the upstream sites may not directly influence the annelids or crustaceans, most of them being associated with habitats with lower water quality (annelids) or brackish water (crustaceans) in the downstream sites. For the communities that had a wide distribution in the LMB (e.g. insects and mollusks), their co-occurrence may be the result of niche expansion or competition, and thus they may have indirect interactions by competing for or facilitating resource availability. For example, an increasing topographical complexity of the streambeds, which can alter the near-bed flow, might enhance feeding success of mussel and suspension-feeding caddisfly communities (Cardinale et al., 2002; Vaughn et al., 2008).

Several studies suggested that, at large spatial scale, the environment is more important than biotic interactions in governing species composition and distribution (Luoto et al., 2006; Pearson and Dawson, 2003), while other suggested both (Araújo and Luoto, 2007). Our finding showed that the combination of component communities had a great influence (adjusted R^2 = 0.84) on the degree of uniqueness in macroinvertebrate community composition (global LCBD indices). However, this is restricted to important factors such as precipitation, land use cover, nutrients and sediment loads, which have been reported to better explain the community composition and distribution of macroinvertebrates (Cai et al., 2012; Nicola et al., 2010), but are not available for our study. Nonetheless, our findings further suggest that the combination of biotic and abiotic conditions explain jointly an appreciable amount of the global LCBD variation (adjusted R^2 = 0.24) and seem functionally important together in governing the uniqueness in community composition, and thus influencing beta diversity and composition of macroinvertebrate communities in the LMB.

5. Conclusion and remarks

The present study revealed the highest number of macroinvertebrate species ever reported from the LMB. The large diversity of different components (annelids, crustaceans, insects and mollusks) led to a great amount of variation, or beta diversity, in overall community composition among studied sites. The important taxa of annelids, crustaceans and mollusks were mostly found in the downstream sites, particularly in the delta, whereas the important taxa of insects were more related to the upstream sites. Most of sites located along the main channels had a high degree of uniqueness in macroinvertebrate taxonomic composition (i.e. high LCBD indices), of which the sites with significant LCBD indices had an exceptionally low richness, which is most likely due to anthropic impacts. An exception was found for one site located in the delta that had a significant LCBD value and moderate macroinvertebrate richness. This is perhaps because of the natural land covers observed on the west side of the river. Mollusk and insect communities, and three environmental variables (water conductivity, river depth and water transparency) were found to be mainly responsible for the variation in LCBD indices.

Our results provide valuable ecological information for selecting locations for conserving different taxonomic groups of macroinvertebrates at broad and small spatial scales. For example, site CM3 and the three other sites (TM1, TM2 and TM4) with significant LCBD indices and low richness are of particular interest for restoration planning, as these locations are experiencing severe degradation of local environments. Site VB4 and the surrounding sites/areas on the west side of the river deserve attention for protection since VB1 had very high richness and VB4 had a significant LCBD index and high richness. The combination of LCBD indices and species richness of the four component communities can thus be used for restoration and conservation planning.

Acknowledgements

The first author is grateful to the Erasmus Mundus LOTUS Unlimited project (European Commission) and the Belmont Forum TLSCC projects for providing the opportunity to do this work. This publication reflects the views only of us the authors, and the European Commission cannot be held responsible for any use that may be made of the information contained therein. We are grateful to the Mekong River Commission who provided the database for our analyzes. Many thanks go to Dr. K.K.Y Wong for his editorial comments on this manuscript. Laboratoire Evolution & Diversité Biologique (EDB) is supported by “Investissement d’Avenir” grants (CEBA, ref. ANR-10-LABX-0025; TULIP, ref. ANR-10-LABX-41).

References

Ahola, L., Mononen, J., Mohaibes, M., 2011. Effects of access to extra cage constructions including a swimming opportunity on the development of stereotypic behaviour in singly housed juvenile farmed mink (Neovison vison). Appl. Anim. Behav. Sci. 134, 201–208.

Al-Shami, S.A., Heino, J., Che Salmah, M.R., Abu Hassan, A., Suhaila, A.H., Madrus, M.R., 2013. Drivers of beta diversity of macroinvertebrate communities in tropical forest streams. Freshw. Biol. 58, 1126–1137.

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., Cornell, H. V., Comita, L.S., Davies, K.F., Harrison, S.P., Kraft, N.J.B., Stegen, J.C., Swenson, N.G., 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol. Lett. 14, 19–28.

Araújo, M.B., Luoto, M., 2007. The importance of biotic interactions for modelling species distributions under climate change. Glob. Ecol. Biogeogr. 16, 743–753.

Arscott, D.B., Tockner, K., Ward, J. V., 2005. Lateral organization of aquatic invertebrates along the corridor of a braided floodplain river. J. North Am. Benthol. Soc. 24, 934–954.

Baird, I.G., Flaherty, M.S., 2005. Mekong River fish conservation zones in southern Laos: assessing effectiveness using local ecological knowledge. Environ. Manage. 36, 439–54.

Boesel, M., 1983. A review of the genus Cricotopus in Ohio, with a key to adults of species of the northeastern United States (Diptera, Chironomidae). Ohio J. Sci. 83, 74–90.

Boyero, L., Ramírez, A., Dudgeon, D., Pearson, R.G., 2009. Are tropical streams really different? J. North Am. Benthol. Soc. 28, 397–403.

Cai, Y., Gong, Z., Qin, B., 2012. Benthic macroinvertebrate community structure in Lake Taihu, China: effects of trophic status, wind-induced disturbance and habitat complexity. J. Great Lakes Res. 38, 39–48.

Cardinale, B.J., Palmer, M.A., Collins, S.L., 2002. Species diversity enhances ecosystem functioning through interspecific facilitation. Nature 415, 426–429.

Clavier, S., Cottet, M., Favriou, P., Phabmixay, S.S., Guédant, P., 2015. Spatial and temporal variation of benthic macroinvertebrates in the Nam Gnom Basin receiving discharged waters from the Nam Theun 2 Reservoir (Lao PDR). Hydroécologie Appliquée 2, 1–27.

Collier, K.J., Lill, A., 2008. Spatial patterns in the composition of shallow-water macroinvertebrate communities of a large New Zealand river. New Zeal. J. Mar. Freshw. Res. 42, 129–141.

Costa, S.S., Melo, A.S., 2008. Beta diversity in stream macroinvertebrate assemblages: among-site and among-microhabitat components. Hydrobiologia 598, 131–138.

Cuong, N.L., Langellotto, G.A., Thuy, T.L., Quynh, V., Thuy, N.T.T., Barrion, A.T., Chen, Y.H., 2016. Arthropod diversity and abundance in wild rice, Oryza rufipogon, in the Mekong Delta, Vietnam. Ann. Entomol. Soc. Am. 109, 542–554.

Dao, H., Kunpradid, T., Vongsambath, C., Do, T., Prum, S., 2010. Report on the 2008 biomonitoring survey of the lower Mekong River and selected tributaries, MRC Technical Paper No. 27. Vientiane, Lao PDR.

Davidson, P.S., Kunpradid, T., Peerapornisal, Y., Nguyen, T.M.L., Pathoumthong, B., Vongsambath, C., Pham, A.D., 2006. Biomonitoring of the lower Mekong River and selected tributaries, MRC Technical Paper No.13. Vientiane, Lao PDR.

Davidson, T.A., Sayer, C.D., Perrow, M., Bramm, M., Jeppesen, E., 2010. The simultaneous inference of zooplanktivorous fish and macrophyte density from sub-fossil cladoceran assemblages: a multivariate regression tree approach. Freshw. Biol. 55, 546–564.

De’ath, G., 2002. Multivariate regression tree: a new technique for modeling species–environment relationships. Ecology 83, 1105–1117.

Dobson, M., Magana, A.E.M., Mathooko, J.M., Ndegwa, F.K., 2002. Detritivores in Kenyan highland streams: more evidence for the paucity of shredders in the tropics? Freshw. Biol. 47, 909–919.

Dray, A.S., Blanchet, G., Borcard, D., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H.H., 2016. Package "adespatial."

Dudgeon, D. (Ed.), 2008. Tropical Stream Ecology. Academic Press, London, UK.

Feld, C.K., Hering, D., 2007. Community structure or function: effects of environmental stress on benthic macroinvertebrates at different spatial scales. Freshw. Biol. 52, 1380–1399.

Friberg, N., Skriver, J., Larsen, S.E., Pedersen, M., Buffagni, A., 2010. Stream macroinvertebrate occurrence along gradients in organic pollution and eutrophication. Freshw. Biol. 55, 1405–1419.

Getwongsa, P., Hanjavanit, C., Sangpradub, N., 2010. Impacts of agricultural land use on stream benthic macroinvertebrates in tributaries of the Mekong River, northeast Thailand. Adv. Environ. Sci. - Int. J. Bioflux Soc. 2, 253–256.

Golfieri, B., Hardersen, S., Maiolini, B., Surian, N., 2016. Odonates as indicators of the ecological integrity of the river corridor: development and application of the Odonate River Index (ORI) in northern Italy. Ecol. Indic. 61, 234–247.

Higgins, C.L., 2010. Patterns of functional and taxonomic organization of stream fishes: inferences based on α, β, and γ diversities. Ecography (Cop.). 33, 678–687.

Hillebrand, H., Blenckner, T., 2002. Regional and local impact on species diversity - from pattern to processes. Oecologia 132, 479–491.

Jost, L., 2007. Partitioning diversity into independent alpha and beta components. Ecology 88, 2427–2439.

Köhler, F., Seddon, M., Bogan, A.E., Tu, D. Van, Sri-aroon, P., Allen, D., 2012. The status and distribution of freshwater molluscs of the Indo-Burma region, in: Allen, D., Smith, K., Darwall, W. (Eds.), The Status and Distribution of Freshwater Biodiversity in Indo-Burma. Gland, Cambridge, pp. 66–89.

Koleff, P., Gaston, K.J., Lennon, J.J., 2003. Measuring beta diversity for presence–absence data. J. Anim. Ecol. 72, 367–382.

Królak, E., Korycińska, M., 2008. Taxonomic composition of macroinvertebrates in the Liwiec River and its tributaries (Central and Eastern Poland) on the basis of chosen physical and chemical parameters of water and season. Polish J. Environ. Stud. 17, 39–50.

Kudthalang, N., Thanee, N., 2010. The assessment of water quality in the upper part of the Chi basin using physicochemical variables and benthic macroinvertebrates. Suranaree J. Sci. Technol. 17, 165–176.

Lamy, T., Legendre, P., Chancerelle, Y., Siu, G., Claudet, J., 2015. Understanding the spatio-temporal response of coral reef fish communities to natural disturbances: insights from beta-diversity decomposition. PLoS One 10, e0138696.

Legendre, P., Borcard, D., Peres-Neto, P., 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecol. Monogr. 75, 435–450.

Legendre, P., De Cáceres, M., 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecol. Lett. 16, 951–963.

Legendre, P., Gallagher, E.D., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280.

Legendre, P., Legendre, L., 2012. Numerical ecology-developments in environmental modelling, 3rd ed. Elsevier Science BV, Amsterdam.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I.-F., He, F., 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90, 663–74.

Legendre, P., Salvat, B., 2015. Thirty-year recovery of mollusc communities after nuclear experimentations on Fangataufa atoll (Tuamotu, French Polynesia). Proc. R. Soc. London B Biol. Sci. 282.

Leigh, C., Sheldon, F., 2009. Hydrological connectivity drives patterns of macroinvertebrate biodiversity in floodplain rivers of the Australian wet/dry tropics. Freshw. Biol. 54, 549–571.

Ligeiro, R., Melo, A.S., Callisto, M., 2010. Spatial scale and the diversity of macroinvertebrates in a Neotropical catchment. Freshw. Biol. 55, 424–435.

Lods-Crozet, B., Castella, E., Cambin, D., Ilg, C., Knispel, S., Mayor-Simeant, H., 2001. Macroinvertebrate community structure in relation to environmental variables in a Swiss glacial stream. Freshw. Biol. 46, 1641–1661.

López-González, C., Presley, S.J., Lozano, A., Stevens, R.D., Higgins, C.L., 2015. Ecological biogeography of Mexican bats: the relative contributions of habitat heterogeneity, beta diversity, and environmental gradients to species richness and composition patterns. Ecography (Cop.). 38, 261–272.

Luoto, M., Heikkinen, R.K., Pöyry, J., Saarinen, K., 2006. Determinants of the biogeographical distribution of butterflies in boreal regions. J. Biogeogr. 33, 1764–1778.

Matthiessen, B., Hillebrand, H., 2006. Dispersal frequency affects local biomass production by controlling local diversity. Ecol. Lett. 9, 652–662.

McCluskey, A., Lalkhen, A.G., 2007. Statistics II: central tendency and spread of data. Contin. Educ. Anaesthesia, Crit. Care Pain 7, 127–130.

Md Rawi, C., Al-Shami, S.A., Madrus, M.R., Ahmad, A.H., 2013. Biological and ecological diversity of aquatic macroinvertebrates in response to hydrological and physicochemical parameters in tropical forest streams of Gunung Tebu, Malaysia: implications for ecohydrological assessment. Ecohydrology 7, 496–507.

Mena, J.L., Vázquez-Domínguez, E., 2005. Species turnover on elevational gradients in small rodents. Glob. Ecol. Biogeogr. 14, 539–547.

Miller, T.E., 1994. Direct and indirect species interactions in an early old-field plant community. Am. Nat. 143, 1007–1025.

Ngor, P., Chhuon, K., Prak, L., 2016. Cambodia completes first pilot study of Tonle Sap mollusc fishery. Catch Cult. 22, 4–13.

Nicola, G.G., Almodóvar, A., Elvira, B., 2010. Effects of environmental factors and predation on benthic communities in headwater streams. Aquat. Sci. 72, 419–429.

Pathoumthong, B., Vongsombath, C., 2007. Macroinvertebrate pilot study for ecological health monitoring in the Lower Mekong Basin, in: Furumai, H., Kurisu, F., Katayama, H., Satoh, H., Ohgaki, S., Thanh, N. (Eds.), Southeast Asian Water Environment 2. IWA Publishing, Cornwall, pp. 123–130.

Pauly, A., 2016. Updating lanternflies biodiversity knowledge in Cambodia (Hemiptera: Fulgoromorpha: Fulgoridae) by optimizing field work surveys with citizen science involvement through Facebook networking and data access in FLOW website. Belgian J. Entomol. 37, 1–16.

Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob. Ecol. Biogeogr. 12, 361–371.

Quang, N., Sinh, N., Tu, N., Lam, P., Lan, N., 2013. Biodiversity of littoral macroinvertebrates in the Mekong River. Tạp chí Khoa học 16–28.

R Core Team, 2013. R: a language and environment for statistical computing.

Ribeiro Jr, P.J., Diggle, P.J., 2015. Package “geoR.”

Rizo-Patrón V., F., Kumar, A., McCoy Colton, M.B., Springer, M., Trama, F.A., 2013. Macroinvertebrate communities as bioindicators of water quality in conventional and organic irrigated rice fields in Guanacaste, Costa Rica. Ecol. Indic. 29, 68–78.

Salmah, M.R.C., Al-Shami, S.A., Abu Hassan, A., Madrus, M.R., Nurul Huda, A., 2014. Distribution of detritivores in tropical forest streams of peninsular Malaysia: role of temperature, canopy cover and altitude variability. Int. J. Biometeorol. 58, 679–690.

Sinclair, C.S., Gresens, S.E., 2008. Discrimination of Cricotopus species (Diptera: Chironomidae) by DNA barcoding. Bull. Entomol. Res. 98, 555–563.

Sodhi, N.S., Koh, L.P., Brook, B.W., Ng, P.K.L., 2004. Southeast Asian biodiversity: an impending disaster. Trends Ecol. Evol. 19, 654–60.

Sor, R., Boets, P., Chea, R., Goethals, P., Lek, S., 2017. Spatial organization of macroinvertebrate assemblages in the Lower Mekong Basin. Limnologica 64, 20–30.

Sor, R., Meas, S., Wong, K.K.Y., Min, M., Segers, H., 2015. Diversity of Monogononta rotifer species among standing waterbodies in northern Cambodia. J. Limnol. 74, 192–204.

Strayer, D.L., Dudgeon, D., 2010. Freshwater biodiversity conservation : recent progress and future challenges. J. North Am. Benthol. Soc. 29, 344–358.

Tonkin, J.D., Stoll, S., Jähnig, S.C., Haase, P., 2015. Variable elements of metacommunity structure across an aquatic-terrestrial ecotone. PeerJ Prepr. 3, e1261.

Valdujo, P.H., Carnaval, A.C.O.Q., Graham, C.H., 2013. Environmental correlates of anuran beta diversity in the Brazilian Cerrado. Ecography (Cop.). 36, 708–717.

Vaughn, C.C., Nichols, S.J., Spooner, D.E., 2008. Community and foodweb ecology of freshwater mussels. J. North Am. Benthol. Soc. 27, 409–423.

Wang, B., Liu, D., Liu, S., Zhang, Y., Lu, D., Wang, L., 2012. Impacts of urbanization on stream habitats and macroinvertebrate communities in the tributaries of Qiangtang River, China. Hydrobiologia 680, 39–51.

Wang, J., Soininen, J., Zhang, Y., Wang, B., Yang, X., Shen, J., 2012. Patterns of elevational beta diversity in micro- and macroorganisms. Glob. Ecol. Biogeogr. 21, 743–750.

Wearn, O.R., Carbone, C., Rowcliffe, J.M., Bernard, H., Ewers, R.M., 2016. Grain-dependent responses of mammalian diversity to land use and the implications for conservation set-aside. Ecol. Appl. 26, 1409–1420.

Wetzel, R.G., 2001. Limnology: lake and river ecosystems, 3rd ed. CA: Academic Press, San Diego.

Whittaker, R., 1972. Evolution and measurement of species diversity. Taxon 21, 213–251.

Whittaker, R., 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol. Monogr. 30, 279–338.

Wootton, J., 1994. Putting the species together: testing the independence of interactions among organisms. Ecology 75, 1544–1551.

Zalinge, N. V., Degen, P., Pongsri, C., Nuov, S., Jensen, J., Hao, N., Choulamany, X., 2003. The Mekong River system, in: Second International Symposium on the Management of Large Rivers for Fisheries Phnom. Phnom Penh, Cambodia, pp. 1–18.

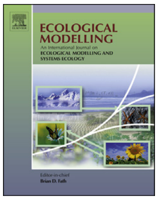
Zalinge, N. V., Thuok, N., 1998. It's big, unique and important: fisheries in the Lower Mekong Basin, as seen from a Cambodian perspective. Catch Cult. 4, 1–8.

Ecological Modelling 354 (2017) 11–19

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Effects of species prevalence on the performance of predictive models

Ratha Sor[a,b,c,*], Young-Seuk Park[d], Pieter Boets[b,e], Peter L.M. Goethals[b], Sovan Lek[a]

[a] *Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse cédex 4, France*
[b] *Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure Links 653, 9000, Ghent, Belgium*
[c] *Department of Biology, Faculty of Science, Royal University of Phnom Penh, Russian Boulevard, 12000, Phnom Penh, Cambodia*
[d] *Department of Life and Nanopharmaceutical Sciences and Department of Biology, Kyung Hee University, Seoul, 130-701, Republic of Korea*
[e] *Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000, Ghent, Belgium*

ARTICLE INFO

Article history:
Received 14 October 2016
Received in revised form 8 March 2017
Accepted 8 March 2017
Available online 21 March 2017

Keywords:
Quadratic effect
Species occurrence
Logistic regression
Random forest
Artificial neural network
Support vector machine
Macroinvertebrates
Habitat suitability
Mekong river

ABSTRACT

Predictive models are useful to support decision making, management and conservation planning. However, the performance of models varies across techniques and is affected by several factors including species prevalence (i.e. the occurrence rate of each species in the total samples). Here, we analysed and compared the performance of four common modelling techniques based on the species prevalence. The occurrence of macroinvertebrates collected at 63 sites along the Lower Mekong Basin was predicted using Logistic Regression, Random Forest, Support Vector Machine and Artificial Neural Network (ANN). Model performance was evaluated using Cohen's Kappa Statistic (Kappa), area under receiver operating characteristic curve (AUC) and error rate. We found a highly significant quadratic effect of species prevalence on the four modelling techniques' performance. Kappa and AUC were less depended on the species prevalence, making them a better measure. The best performance (Kappa and AUC) was reached when predicting species with an intermediate prevalence (e.g. 0.4–0.6). The four modelling techniques significantly yielded different performances ($p < 0.01$), of which ANN performed generally better when using the complete prevalence range (i.e. 0.0–1.0) and the lower prevalence range (i.e. <0.1). However, the four techniques similarly performed when predicting species with a higher prevalence range (i.e. ≥0.3). Our results provide useful insights into the application of modelling techniques in predicting species occurrence and how their performance varies for species with different prevalence ranges. We suggest that the selection of appropriate modelling techniques should carefully take into account the species prevalence, particularly in the case of rare and generalist species.

1. Introduction

Various modelling techniques have been widely implemented in different ecological systems, e.g. terrestrial, freshwater lentic and lotic, and marine ecosystems (Guo et al., 2015; Lek et al., 1996; Lencioni et al., 2007; Park et al., 2003; Schröder et al., 2007). The techniques applied are generally used to investigate or to explain the relationship between the occurrence or abundance of studied species and environmental variables or to predict the relationships being measured (Boets et al., 2013; Goethals et al., 2007). The use of modelling techniques to combine both explaining and predicting such relationships is also commonly applied (Call et al., 2016; Roura-Pascual et al., 2009).

The performance of data-driven predictive models is affected by several factors including species prevalence (Brotons et al., 2004; Hernandez et al., 2006; Stokland et al., 2011). In most cases, models predicting species which have unequal occupied and unoccupied samples/sites result in a low performance. With a species having a high prevalence, models tend to become better at predicting the presence of that species, and vice-versa for less occurring species (McPherson and Jetz, 2007). Both cases consequently lead to a low model performance when considering the correct prediction of both the presence and absence of a species. Moreover, it has been demonstrated that species prevalence affects the performance of models in a nonlinear way. For example, Guo et al. (2015) and Manel et al. (2001) reported the nonlinear effect of species prevalence on the performance of models predicting the occurrence of fish and macroinvertebrates. A similar finding has also been revealed for

* Corresponding author at: Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse cédex 4, France.
E-mail address: sorsim.ratha@gmail.com (R. Sor).

http://dx.doi.org/10.1016/j.ecolmodel.2017.03.006

models predicting the distribution of plants and birds (Allouche et al., 2006; McPherson et al., 2004).

Applications of predictive models have provided knowledge and understanding of the ecology and behaviour of studied taxa, which could support decision making, management and conservation planning. For instance, Chen et al. (2015) used different predictive models as an assessment approach to explain and predict the success of invasive species in China. In addition to the increased use of predictive models, an ensemble modelling framework is recommended when aiming to identify important factors influencing model performance (Araújo and New, 2007). With the ensemble modelling approach, some modelling techniques such as Random Forest and Artificial Neural Networks are found to yield a better predictive performance (Grenouillet et al., 2011; Guo et al., 2015; Segurado and Araujo, 2004). However, although there have been studies assessing the performance of predicting models from an ensemble modelling framework, many have not considered analysing the performance based on a complete prevalence range nor comparing the performance based on different prevalence ranges.

The Lower Mekong Basin (LMB) which is known for its high biodiversity (Sodhi et al., 2004) is a breeding ground of numerous endemic, threatened and endangered species of fish, molluscs and crustaceans (Davidson et al., 2006; Zalinge and Van Thuok, 1998). Therefore, it is useful to get more insight into this region based on predictive models which are applicable for different taxonomic groups inhabiting this particular area. To date, the data covering a large spatial scale of the LMB is only available for fish and macroinvertebrates, which were collected by the Mekong River Commission (MRC). The fish data were collected only from the main channel (Poulsen and Viravong, 2001), while macroinvertebrates were collected from both the tributaries and the main channel (Dao et al., 2010). In this study, we used the macroinvertebrate data, sampled over 5 successive years (2004–2008), to build predictive models, which can provide insights on a wide range of keystone species occupying the LMB as well as the neighbouring regions.

The objectives of the present study are to utilize different modelling techniques to 1) predict the occurrence of macroinvertebrate species in the LMB and analyse how the species prevalence (i.e. the occurrence rate of each species in the total samples) affects the behaviour of modelling techniques' performance, and 2) compare the performance of the applied techniques based on the complete prevalence range (i.e. 0.0–1.0), and based on different prevalence ranges (i.e. at a 0.1 interval).

2. Methods

2.1. Data collection and processing

Benthic macroinvertebrates were sampled at 63 sampling sites along the main channel of the LMB and its tributaries by the MRC. This sampling was carried out once a year in March during the dry season from 2004 to 2008. To obtain as much information as possible on macroinvertebrates inhabiting the main river and the tributaries, the MRC collected samples at three locations from the benthic zone of each sampling site: near the left and right banks, and in the middle of the rivers. At each location, a minimum of three samples (where inter-sample variability is low, e.g. tributaries) to a maximum of five samples (where inter-sample variability is higher, e.g. the main channel and the delta) were collected using a Petersen grab sampler. With the grab which has a sampling area of 0.025 m^2, four sub-samples were taken and pooled to give a single sample covering a total area of 0.1 m^2. In total, between nine (3 samples × 3 locations) and fifteen (5 samples × 3 locations) pooled samples were collected at each sampling site. Each pooled sample was rinsed using a sieve (0.3 mm mesh size). In the field, samples were sorted and then preserved by adding 10% formaldehyde to obtain a final concentration of about 5%. In the laboratory, the samples were identified to the lowest level possible and counted using a compound microscope (40–1200 magnification) or a dissecting microscope (16–56 magnification). The abundance data of macroinvertebrates per sample (a total area of 0.1 m^2) was averaged across all samples (between 9 and 15 samples) collected from each sampling site.

At the sampling site, geographical coordinates and altitude were determined with a GPS (Garmin GPS 12XL). All physical-chemical variables were measured at the three locations where macroinvertebrates were sampled. River width was measured in the field using a Newcon Optik LRB 7 × 50 laser rangefinder, and the river depth was measured using a line metre. Water temperature, dissolved oxygen, pH and water conductivity were measured using a handheld water quality probe (YSI 556MP5). To get a more reliable determination of each variable, the measurement reading was taken at the surface (0.1–0.5 m) and at a depth of 3.5 m or at a maximum depth of the river (wherever less than 3.5 m) and then the average value was recorded for each location. Water transparency was measured with a Secchi disc by lowering it into the water and recording the depth at which it was no longer visible (Dao et al., 2010). The recorded data of each physical-chemical variable was based on the averaged value across the three sampling locations of each site. The distance from the sea was measured by drawing a line from the sea to each locality using GIS-software (ArcGIS version 10.0).

A total of 108 samples were collected from the 63 sampling sites (Fig. 1). Because of unequal sampling efforts (i.e. unequal and different number of samples at each site during the 5-year sampling period) and missing values of environmental variables, we used median values from the collected data to represent each site in our analyses, as suggested by McCluskey and Lalkhen (2007). Therefore, 63 samples remained for the analyses. In total, 299 taxa were obtained from the dataset, of which 131 taxa were insects, 98 were molluscs, 38 were crustaceans and 32 were annelids. The most commonly identified insects belonged to Diptera (37 taxa), Ephemeroptera (32), Odonata (22) and Trichoptera (20). For molluscs, Caenogastropoda (50 taxa), Unionida (18) and Veneroida (15) were represented the most. Most crustaceans belonged to Palaemonidae (10 taxa) and Corophiidae (6 taxa), while most annelids belonged to Naididae (15 taxa) and Nereididae (5 taxa). The detailed information of taxonomic resolution is provided in the Supplementary data Appendix A.

The abundance data of macroinvertebrates from the 63 sites were converted to presence-absence data to analyse how species prevalence (presence/absence) affects the performance of predictive models. Species prevalence was defined as the occurrence rate of each species in the total samples. The species prevalence of a species is an index ranging from 0 to 1, indicating the lowest to highest occurrence rate of that species over all samples. The obtained prevalence values from all macroinvertebrate species formed a complete prevalence range for the present study. For a later analysis, the complete prevalence range was grouped into different ranges based on an interval of 0.1. In other words, the species having a prevalence value between 0.0 and 0.1 were aggregated in a group, and the species having a prevalence between 0.1 and 0.2 were aggregated in another group, and so forth (see Appendix A in Supplementary material).

2.2. Predictions

In our predictive models, we used the presence/absence of each species as the response variable. The measured environmental variables used as the input predictors were: altitude, river width, river

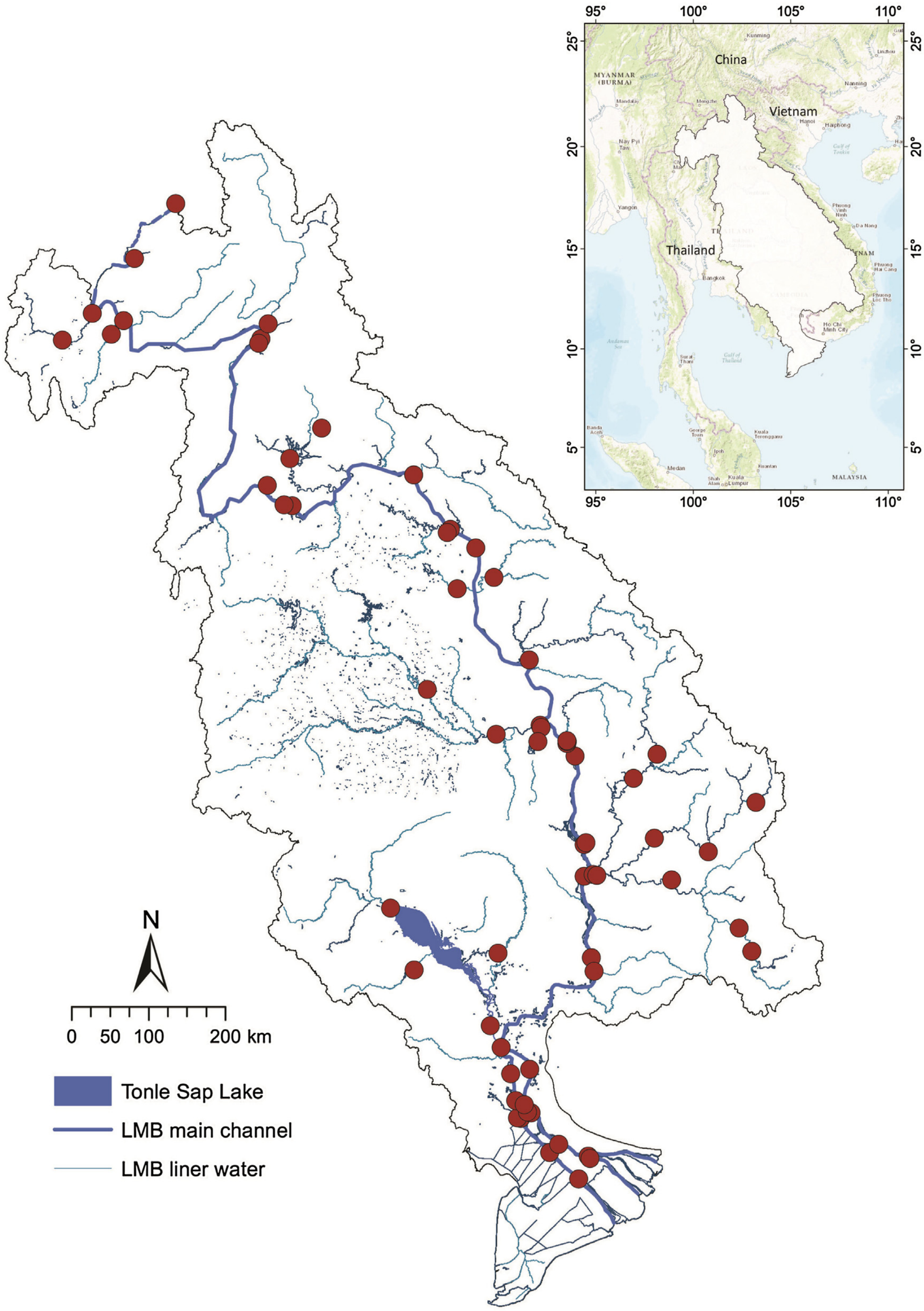


Fig. 1. The Lower Mekong Basin (LMB) and macroinvertebrate sampling locations (shaded dots).

depth, distance from the sea, water temperature, dissolved oxygen, water conductivity, pH and Secchi depth. The summary of all environmental variables is provided in Table 1. Before building the models, the data of these predictor variables were normalized using the zero minimum (Fig. 2). This normalization approach is appropriate for variables that highly vary (different measuring scales). The normalized values were based on the variables' standard deviation, and the range of all variables remains constant at the same time (Dębkowska and Jarocka, 2013). For every prediction, species that occurred only in one instance (a prevalence of ~0.02) were

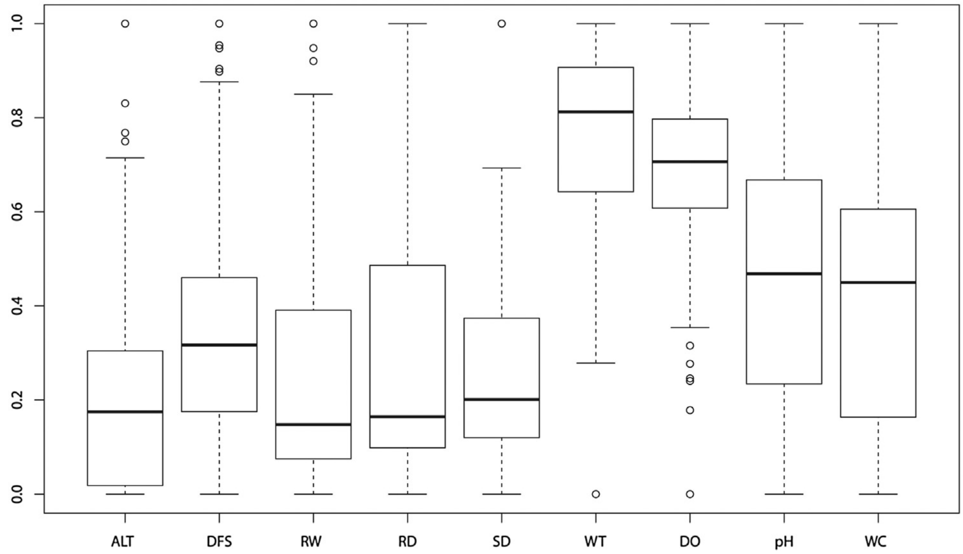


Fig. 2. Box and whisker plots of normalized environmental variables. Rectangles delineate the first and third quartiles, dark bars are the medians, the lower and upper bars are the minima and maxima, and the circles are outliers. ALT: altitude, DFS: distance from the sea, RW: river width, RD: river depth, SD: Secchi depth, WT: water temperature, DO: dissolved oxygen, WC: water conductivity.

Table 1
Minimum, maximum and mean values and standard deviation for environmental predictors across the 63 sampling sites.

Predictor	Unit	Min	Max	Mean	Standard deviation
Altitude	m	3	546	127	132
Water temperature	°C	16.7	31	27.4	3
Dissolved oxygen	mg/L	2.7	9.3	7.1	1.3
Water conductivity	mS/m	3.9	66.6	17.9	11.9
Distance from the sea	km	82	2597	998	686
River width	m	11	1629	466	466
River depth	m	0.4	15	4.8	3.9
Secchi depth	m	0.2	3	0.9	0.6
pH	–	6.8	8.4	7.6	0.6

removed from the data. We removed these species because we used Leave-One-Out cross-validation (LOO), due to our small sample size, to validate the models. With LOO, it is not feasible to split data into training and validation sets for species that have only one occurrence instance.

2.2.1. Model selection and validation

The occurrence of macroinvertebrate species was predicted using four modelling techniques which are commonly used in ecology, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN). LR is an extension of the general linear model or so-called generalized linear model (GLM, McCullagh and Nelder, 1989). In LR, a binary response (presence/absence) and a “logit” link function are used. The “logit” link function systematically defines the relationship between the mean value of each response variable and the systematic part of the model, which is a function of the predictor variables. Values between 0 and 1 are the outputs for LR. A threshold value of 0.5 was used to indicate if a species was present or not. If the output value was >0.5, it indicated the occurrence (the presence) of the macroinvertebrate species and vice versa.

RF is a classification method that grows an ensemble of 500 trees. Each tree casts a unit vote for the most popular class according to the input variables. The growing process of RF is by bootstrap aggregating (or bagging), where a tree is randomly grown from the dataset, and this process can give substantial gains in the accuracy of predicting models (Breiman, 1996), and it requires no pruning. The random split selection of the bagging process provides nodes of the tree where the split is selected using a subset of the input predictors that are chosen for each tree.

SVM is a supervised machine learning approach that is based on a linear classifier using a maximum hyperplane to separate two classes (Vapnik, 1995). For nonlinear classification, SVM uses kernel functions implicitly to map training data into a higher-dimensional feature space and to compute separating hyperplanes in a way to maximize the margin between classes. The maximum separation hyperplane is defined by a set of support vectors, which are a function of the training data that lie on/the closest to the separating margin. For each class, there is always at least one support vector present. The radial basis kernel function was applied for all SVM models.

ANN is a non-linear statistical data modelling tool used for prediction and classification. ANN is based on the basic function of biological neural networks and on a set of linked computed neurons. One of the most popular approaches of neural networks is a multilayer feedforward neural network (Goethals et al., 2007; Lek et al., 1996; Lek and Guégan, 1999). This is a backpropagation network that trains the data using a backpropagation algorithm. This network comprises three layers: an input layer, one or more hidden layers and an output layer. Neurons from one layer are in one direction linked to all neurons in the subsequent layers. In this way, connection weights among the neurons are calculated during the training phase. For detailed information on the above mentioned techniques as well as for practical examples of the used techniques, we refer to Van Echelpoel et al. (2015).

To validate the models built using the four modelling techniques, Leave-One-Out cross-validation (LOO) was used. LOO is one of the cross-validation methods using each observation as the validation when the model is trained with the remaining N-1 observations (where N is the total number of observations). It is a time-consuming approach as at a time, each observation is temporarily removed from the dataset, and this process is repeated until every single observation is rotationally left out. LOO has an advantage when N is small (as is the case in this study) because this validation method can provide a nearly unbiased estimation of the accuracy (Cawley and Talbot, 2003; Efron, 1983).

All modelling procedures were performed in the R language program (R Core Team, 2013). LR was performed using the *glm()* function of the *stats* package (R Core Team, 2013), RF using the *ran-*

domForest() function of the *randomForest* package (Breiman, 2001), ANN using the *nnet()* function of the *nnet* package (Ripley and Venables, 2016) and SVM using the *svm()* function of the *e1071* package (Dimitriadou et al., 2009).

2.2.2. Model performance measures and statistical analyses

To evaluate the four modelling techniques, three types of performance measures were used: Cohen's Kappa Statistic (Kappa), area under receiver operating characteristic curve (AUC) and error rate (ER) of the prediction. AUC was calculated as the quantitative principal performance metric. The ER (i.e. misclassification rate) was estimated measuring the fraction of all instances that were not correctly predicted. Each performance measure of the four modelling techniques was regressed against the species prevalence using two types of models: a linear and a quadratic model.

To examine overall performance differences among the four modelling techniques, we performed a Friedman test (nonnormal distribution). Mean performance values and standard errors were compared to identify which technique better performed based on the complete prevalence and based on different prevalence ranges. P-values ≤0.05 were considered to indicate significant differences. The Friedman test was performed using the *friedman.test()* function of the *stats* package in R (R Core Team, 2013).

3. Results

3.1. Effect of species prevalence

The species prevalence ranged from 0.03 to 0.73. Linear regression models showed that the prevalence always (for each modelling technique) had a positive and significant effect on the three performance measures used (Table 2). The proportion of variance of Kappa and AUC explained by the linear models ranged from 0.02 to 0.15, while the proportion of variance of error rate (ER) explained ranged from 0.59 to 0.63. When the same data were analysed using a quadratic regression model, the explained proportion of variance of Kappa and AUC increased from 0.06 to 0.31, and the explained proportion of variance of ER increased from 0.66 to 0.84. The coefficient of the quadratic term was always negative and highly significant ($p < 0.001$, Table 2).

Based on Kappa and AUC, the four modelling techniques had a low performance when predicting species with a prevalence <0.1 and ≥0.6. A high performance was observed when predicting species with an intermediate prevalence between 0.4 and 0.6. Based on ER, the prediction error of the models increased from the lowest to the highest prevalence (Fig. 3).

3.2. Model performance

The overall performance among the four modelling techniques was significantly different when considering the three performance measures: Kappa (Friedman chi-squared, F = 12.3; p = 0.006), AUC (F = 11.2, p = 0.01) and ER (F = 350, p < 0.001). The highest mean Kappa (0.19) and mean AUC (0.60) were obtained for ANN, followed by LR (mean Kappa: 0.16, mean AUC: 0.59) and RF (mean Kappa: 0.12, mean AUC: 0.55), while SVM yielded the lowest mean Kappa (0.06) and mean AUC (0.53). On the other hand, a lower mean ER (0.09) was obtained for RF and SVM, while ANN and LR had a higher mean ER (0.13 and 0.16, respectively).

Based on Kappa and AUC, the performance of the models varied for different prevalence ranges (Fig. 4). ANN and LR performed better than RF and SVM for the prevalence range <0.1; ANN, RF and LR performed better than SVM for the prevalence range 0.1-0.2; ANN and RF performed better than LR and SVM for the prevalence range 0.2–0.3, but not significantly different according to the standard error. Based on ER, RF and SVM performed better than ANN and LR in predicting species with a prevalence range <0.1, between 0.1 and 0.2 and between 0.2 and 0.3. Based on all calculations, the model performance was not significantly different for the prevalence range ≥0.3 (Fig. 4).

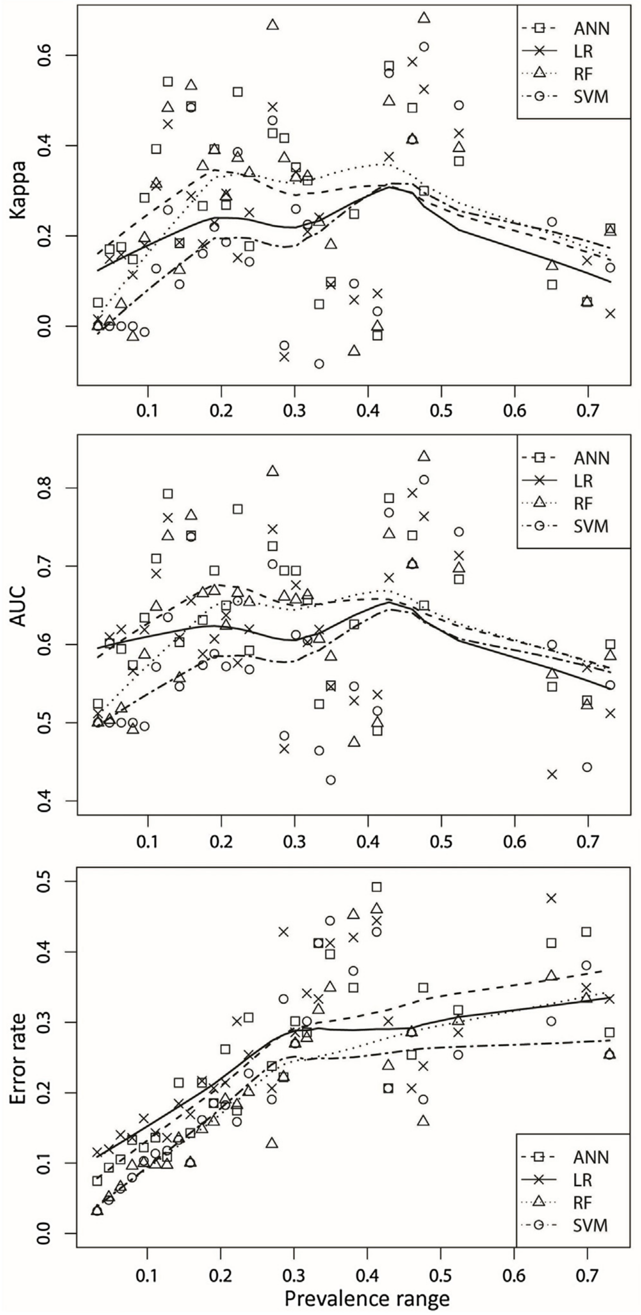


Fig. 3. Effects of species prevalence on the performance measures of the predictive models. Each symbol is the mean performance value of the corresponding modelling techniques. The lines are LOWESS smoothers. Kappa: Cohen's Kappa Statistic, AUC: area under the curve, LR: logistic regression, ANN: artificial neural network, RF: random forest, SVM: support vector machine.

4. Discussion

4.1. The effect of species prevalence

We found a highly significant quadratic effect of species prevalence on the performance of the four modelling techniques for all three measures, which clarifies the nonlinear relationship found

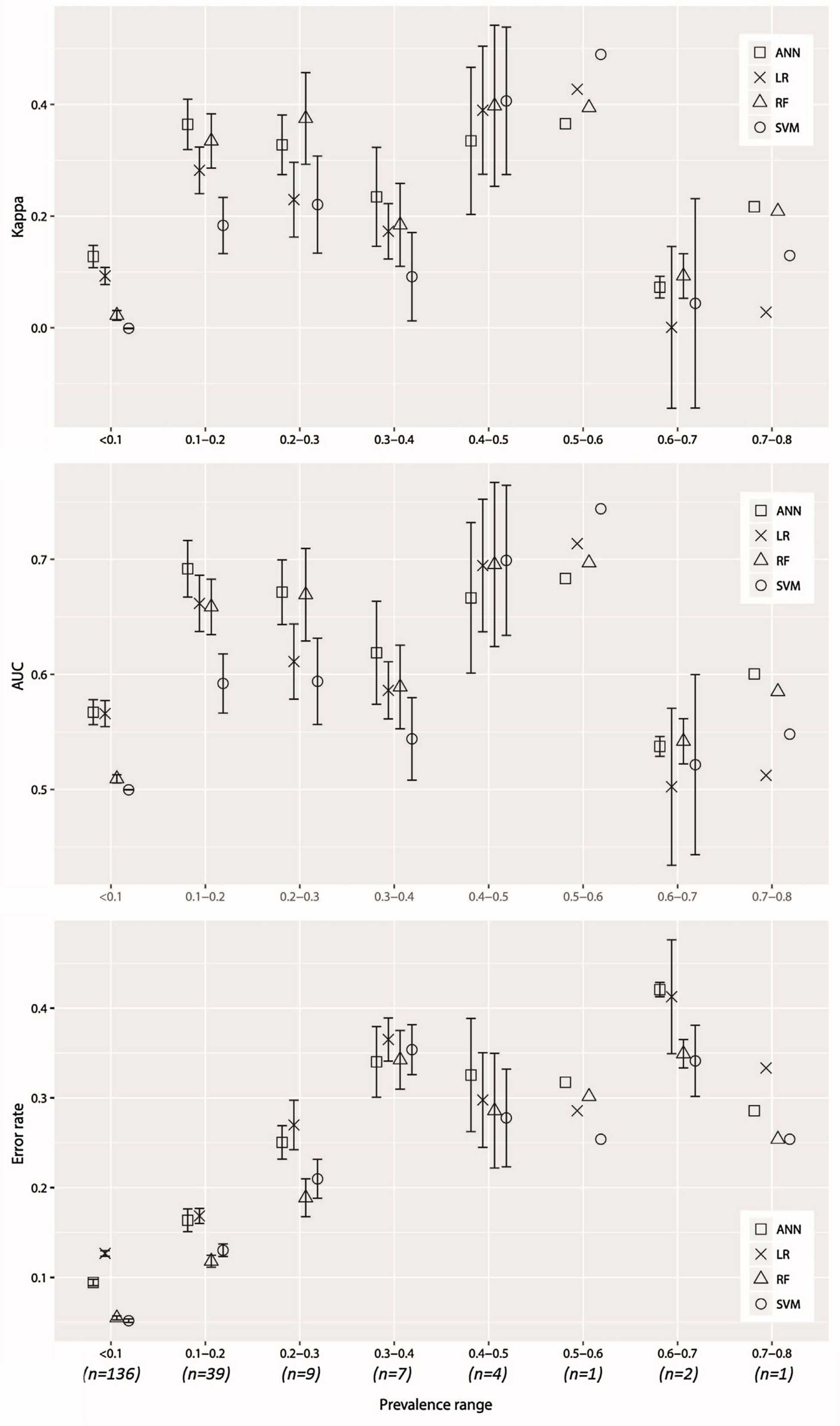


Fig. 4. Mean values and standard error bars showing the performance measures of each modelling technique based on different species prevalence ranges. Kappa: Cohen's Kappa Statistic, AUC: area under the curve, LR: logistic regression, ANN: artificial neural network, RF: random forest, SVM: support vector machine, *n*: the number of predicted species corresponding to each prevalence range.

Table 2
Results of linear regression models ($y = \alpha + \beta_1 x$) and quadratic regression models ($y = \alpha + \beta_1 x + \beta_2 x^2$) for the effects of the prevalence of macroinvertebrate species on the three performance measures (Kappa, AUC and ER) of the modelling techniques. Kappa: Cohen's Kappa Statistic, AUC: area under the curve, ER: error rate, LR: logistic regression, RF: random forest, ANN: artificial neural network, SVM: support vector machine. Asterisks indicate significance levels of regression coefficients.

Measure	Regression Model	β_1	β_2	Adjusted R^2
LR				
Kappa	Linear	0.006*		0.04
	Quadratic	0.613**	−0.91***	0.12
AUC	Linear	0.003*		0.02
	Quadratic	0.234	−0.46***	0.06
ER	Linear	0.008***		0.61
	Quadratic	0.861***	−0.23***	0.66
RF				
Kappa	Linear	0.012***		0.15
	Quadratic	1.226***	−1.29***	0.31
AUC	Linear	0.005***		0.15
	Quadratic	0.582***	−0.61***	0.31
ER	Linear	0.009***		0.73
	Quadratic	1.008***	−0.34***	0.81
SVM				
Kappa	Linear	0.009***		0.12
	Quadratic	0.927***	−0.72***	0.19
AUC	Linear	0.004***		0.11
	Quadratic	0.347***	−0.35***	0.18
ER	Linear	0.009***		0.73
	Quadratic	1.032***	−0.40***	0.84
ANN				
Kappa	Linear	0.007**		0.04
	Quadratic	0.736**	−1.07***	0.12
AUC	Linear	0.003**		0.03
	Quadratic	0.357**	−0.56***	0.11
ER	Linear	0.009***		0.59
	Quadratic	0.993***	−0.31***	0.66

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

in previous studies. Species prevalence has been reported to influence model performance in a nonlinear way. This finding has been demonstrated from the prediction of freshwater aquatic species, e.g. fish and macroinvertebrates (Guo et al., 2015; Manel et al., 2001) and of terrestrial plant and bird species (Allouche et al., 2006; McPherson et al., 2004). Moreover, our results also indicated that the performance of predictive models responds in a positive linear way as long as the prevalence remains below 0.2 for Kappa and AUC or below 0.3 for ER (Fig. 3). At a higher prevalence than the intermediate range (e.g. 0.4–0.6), the performance of the models, based on Kappa and AUC, showed a negative relationship. Although we did not have many samples with a high prevalence to support this assumption, evidence that a higher prevalence than the intermediate negatively affects the performance was found from previous studies (e.g. Guisan and Hofer, 2003; McPherson and Jetz, 2007; Stockwell and Peterson, 2002). However, the negative relation between the higher prevalence and the ER was not observed (Fig. 3). Nevertheless, when considering the whole prevalence range for all the three performance measures, our findings suggest rather a highly significant quadratic than a simple linear effect, as reflected by a higher proportion of variance explained by the quadratic regression models (Table 2).

Quadratic effects of species prevalence have been shown to influence Kappa and ER, but not AUC (Allouche et al., 2006; Manel et al., 2001). Our findings also indicate a quadratic effect of species prevalence on AUC. This notable difference may be due to the input predictor variables, and the type of response variables (e.g. fauna vs. flora). Geomorphological variables (e.g. river depth, river width and distance from the sea) and water chemistry (e.g. dissolved oxygen, water conductivity and pH) have shown to strongly influence the occurrence and distribution of macroinvertebrates (Chadwick et al., 2006; Rizo-Patrón et al., 2013), whereas climatic variables and vegetation density are known to have important effects on plant and bird distributions, respectively (Amissah et al., 2014; Toledo et al., 2012). For example, Allouche et al. (2006) found no significant quadratic effect of prevalence on AUC when modelling the distribution of woody plants using three climatic variables (i.e. a mean value of annual rainfall, of daily temperature and of minimum temperature). On the other hand, McPherson et al. (2004) found to some extent a quadratic dependency of AUC on the prevalence when using land surface and air temperature, vapour pressure deficit and normalized difference in vegetation index to predict the distribution of birds. This indicates that different predictor variables used to predict the same type of response variable or vice-versa are more likely to yield a different model performance, and thus affect the relationship between species prevalence and model performance when they are regressed against each other.

For all models, the best performance (i.e. Kappa and AUC) was found for species with an intermediate prevalence (e.g. 0.4–0.6, Figs. 3 and 4). In between this prevalence range, there is a smaller bias for the models to select presence/absence data for training and validation sets. A more or less equal distribution between the presence and absence, which are likely to result in many correctly predicted instances of both the true positive and true negative fractions, could be the main reason responsible for the best predictive performance (Allouche et al., 2006; Manel et al., 2001).

For very common (prevalence ≥0.6) and for rare species (prevalence <0.1, see Appendix A in Supplementary material), the models yielded a low performance (i.e. Kappa and AUC). In these cases, there is a high imbalance between the presence and absence data. Therefore, the models can correctly predict many true positive instances and a few or perhaps no true negative instances for common species, and vice versa for rare species. This explains the low performance of the models because Kappa and AUC are designed to reflect model performance in absence and presence instances simultaneously (Cicchetti and Feinstein, 1990; Cohen, 1960 Zweig and Campbell, 1993). Thus, a few or no instances of either the true positive or the true negative fraction results in a low performance. On the other hand, opposite results were found for rare species when the performance was based on the error rate (ER); the predicting models had a highly reliable performance (a low prediction error) (Figs. 3 and 4). Based on previous research, ER is considered to be a misleading measure (Fielding and Bell, 1997; Manel et al., 1999, 2001). One main reason is that ER only takes into account misclassified instances to estimate the error rate. Indeed in our results, the ER does not reflect the explanation that rare species are well predicted. This suggests that ER is giving an ambiguous performance of models for predicting the presence of species with a lower occurrence or of rare species.

4.2. *Comparison of modelling techniques' performance*

Overall, based on Kappa and AUC, we found that ANN performed the best across the complete prevalence range, whereas SVM had the lowest reliability, and RF and LR had an intermediate performance. While some authors (e.g. Mastrorillo et al., 1997; Pearson et al., 2002; Segurado and Araujo, 2004) agree that ANN provides advantages over other techniques for predicting species occurrence, others found RF to be better (Gallardo and Aldridge, 2013; Grenouillet et al., 2011). Studies comparing model performance usually suggest different modelling techniques based on the performance measures. For example, Guo et al. (2015) compared the performance of nine modelling techniques and found no significant difference, but suggested RF as a better technique, especially in terms of interpretation. Whereas Segurado and Araujo (2004) compared nine modelling techniques' performance and found a significant difference and suggested ANN to perform better. However, both studies may not be fully comparable due to different criteria in selecting different environmental predictors.

On the other hand, our results, based on the overall mean ER, indicated that SVM and RF performed the best, as reflected by the lowest mean error rate (0.09 ± 0.01), whereas LR had the highest error rate (0.19 ± 0.01). The ER was highly dependent on the species prevalence, as implied by a highly explained proportion of variance of the ER (Table 2). Therefore, our findings, along with the recommendation from previous studies (e.g. Fielding and Bell, 1997; Manel et al., 2001, 1999), suggest that ER is highly biased, compared to Kappa and AUC, to evaluate model performance.

The best model in terms of performance may not be the best model in terms of applicability. The performance is more likely related to theoretical knowledge of how each model works and to the parameterization settings (Araújo and Guisan, 2006; Elith and Graham, 2009). Eliminating less important variables or increasing the number of tree nodes when fitting a model may result in a higher model performance. However, these settings have to be carefully taken into account because they may lead to an overfitted model, which produces a less general result with a lower applicability to different situations (Babyak, 2004). Model applicability is dependent on the type of data (e.g. missing values and data distribution, De'ath and Fabricius, 2000; Therneau and Atkinson, 1997) and on where the data is derived (e.g. different ecological regions, Guisan and Thuiller, 2005; Randin et al., 2006). Most of the studies evaluate and apply modelling techniques based on the data collected within particular regions, but the applicability of those techniques to a different geographical range is hardly assessed (Fielding and Haworth, 1995; Kleyer, 2002; Özesmi and Mitsch, 1997). As such, our findings may reveal a unique behaviour of each modelling technique and its applicability for the Lower Mekong Basin and the neighbouring areas rather than for other geographical areas. This is because the performance of the same technique applied in different geographical regions can be variable (Randin et al., 2006). Accordingly, it is important to balance the model performance and its applicability when comparing modelling techniques.

Furthermore, each technique may have its own distinct characteristics regarding adjustment to the response variables (Guisan and Zimmermann, 2000). Based on Kappa and AUC, our study found that ANN and LR were more suitable to predict the occurrence of rare species (e.g. prevalence range <0.1); although the performance was not very high, it was still higher than the performance of RF and SVM. Rare species mostly occur in a particular geographical region and prefer a specific set of environmental conditions (Prendergast et al., 1993). For example, Ephemeroptera (15 taxa which made up 31% of all insect species and 11% of all species that have a prevalence range <0.1) and Trichoptera (5 taxa, see Appendix A), two groups of sensitive species, mostly occur in mountainous areas with clean or unpolluted environments (Md Rawi et al., 2013; Suhaila and Che Salmah, 2014). Some species of molluscs (e.g. *Gyraulus* and *Thiara*) have been recorded from restricted habitats (Choubisa and Sheikh, 2013). Many annelid species in the family Naididae and crustacean species in the family Palaemonidae have been mainly found from a more polluted environment and from estuaries or brackish water, respectively (De Grave et al., 2008; Martins et al., 2008). Perhaps, these site-specific environmental conditions could be well predicted by models using ANN and LR. The performance of the same modelling techniques (ANN and LR) and of RF were comparable when predicting species with a prevalence range between 0.1 and 0.2 (Fig. 4). For species having a wider occupancy (e.g. dipteran insects and a mollusc species *Corbicula tenuis*), the performance among the modelling techniques was not significantly different. This could be due to the fact that those widespread species (e.g. species with an intermediate prevalence or higher) are able to respond to a wide range of environmental conditions, which could be predicted by each type of technique used. Therefore, our findings suggest that species prevalence should be carefully taken into account when assessing model performance and predicting species occurrences. Selecting modelling techniques to predict a given species should also depend on the characteristics of the data and the purpose of the prediction.

5. Conclusion and remarks

Overall, we found a highly significant quadratic effect of species prevalence on the performance of the four modelling techniques. Compared to error rate (ER), the dependency of Kappa and AUC on the species prevalence was rather low, making them a better measure of model performance. A maximum performance was obtained when the species prevalence range was situated between 0.4 and 0.6. ANN generally provided a better overall performance than other modelling techniques and yielded a higher reliability when predicting species with a low occurrence. Our findings could offer useful knowledge regarding the understanding and possible proposition of modelling techniques for other species of interest that inhabit the Lower Mekong Basin and the neighbouring areas.

This study provides clear insights into the application of different modelling techniques when predicting species occurrence and how their performances vary for different species prevalence ranges. Each modelling technique has its strengths and weaknesses. Thus the selection of an appropriate technique should depend on data availability and the purpose of the study, and should balance between model performance and applicability. Modelers may consider a technique (e.g. ANN) that is seen to be generally robust across all species and across species with a small distribution range or a low occurrence (e.g. prevalence range <0.1). Environmental predictors and species prevalence (as shown in this paper), should be carefully taken into account for studies attempting to assess the distribution of a given species.

Acknowledgements

The first author would like to sincerely thank Erasmus Mundus LOTUS project (European Commission) and the Forum Belmont TLSCC project for providing the possibility to do this research. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. We are grateful to the Mekong River Commission who provided the database for our analyses.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2017.03.006.

References

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232.

Amissah, L., Mohren, G.M.J., Bongers, F., Hawthorne, W.D., Poorter, L., 2014. Rainfall and temperature affect tree species distribution in Ghana. J. Trop. Ecol. 30, 435–446.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47.

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 66, 411–421.

Boets, P., Lock, K., Goethals, P.L.M., 2013. Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. Ecol. Inform. 17, 73–81.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 1–35.

Brotons, L., Thuiller, W., Araújo, M.B., Hirzel, A.H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography (Cop.) 27, 437–448.

Call, A., Sun, Y.X., Yu, Y., Pearman, P.B., Thomas, D.T., Trigiano, R.N., Carbone, I., Xiang, Q.Y., 2016. Genetic structure and post-glacial expansion of *Cornus florida* L. (Cornaceae): integrative evidence from phylogeography population demographic history, and species distribution modeling. J. Syst. Evol. 54, 136–151.

Cawley, G.C., Talbot, N.L.C., 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognit. 36, 2585–2592.

Chadwick, M.A., Dobberfuhl, D.R., Benke, A.C., Huryn, A.D., Suberkropp, K., Thiele, J.E., 2006. Urbanization affects stream ecosystem function by altering hydrology, chemistry, and biotic richness. Ecol. Appl. 16, 1796–1807.

Chen, L., Peng, S., Yang, B., 2015. Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. Biol. Invasions 17, 2187–2198.

Choubisa, S.L., Sheikh, Z., 2013. Freshwater snails (Mollusca: Gastropoda) as bio-indicators for diverse ecological aquatic habitats. Cibtech J. Zool. 2, 22–26.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low Kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 543–549.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Dębkowska, K., Jarocka, M., 2013. The impact of the methods of the data normalization on the result of linear ordering. Acta Univ. Lodz. – Folia Oeconomica 286, 181–188.

Dao, H., Kunpradid, T., Vongsambath, C., Do, T., Prum, S., 2010. Report on the 2008 Biomonitoring Survey of the Lower Mekong River and Selected Tributaries, MRC Technical Paper No. 27, Vientiane, Lao PDR.

Davidson, P.S., Kunpradid, T., Peeraporrnisal, Y., Nguyen, T.M.L., Pathoumthong, B., Vongsambath, C., Pham, A.D., 2006. Biomonitoring of the lower Mekong River and Selected Tributaries, MRC Technical Paper No.13, Vientiane, Lao PDR.

De Grave, S., Cai, Y., Anker, A., 2008. Global diversity of shrimps (Crustacea: Decapoda: Caridea) in freshwater. Hydrobiologia 595, 287–293.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., Leisch, M.F., 2009. Package e1071.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331.

Elith, J., Graham, C.H., 2009. Do they? How do they? Why do they differ? on finding reasons for differing performances of species distribution models. Ecography (Cop.) 32, 66–77.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Fielding, A.H., Haworth, P.F., 1995. Testing the generality of bird-habitat models. Conserv. Biol. 9, 1466–1481.

Gallardo, B., Aldridge, D.C., 2013. Evaluating the combined threat of climate change and biological invasions on endangered species. Biol. Conserv. 160, 225–233.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat. Ecol. 41, 491–508.

Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography (Cop.) 34, 9–17.

Guisan, A., Hofer, U., 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography. J. Biogeogr. 30, 1233–1243.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Guo, C., Lek, S., Ye, S., Li, W., Liu, J., Li, Z., 2015. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. Ecol. Model. 306, 67–75.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography (Cop.) 29, 773–785.

Kleyer, M., 2002. Validation of plant functional types across two contrasting landscapes. J. Veg. Sci. 13, 167–178.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Lencioni, V., Maiolini, B., Marziali, L., Lek, S., Rossaro, B., 2007. Macroinvertebrate assemblages in glacial stream systems: a comparison of linear multivariate methods with artificial neural networks. Ecol. Model. 203, 119–131.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J. Appl. Ecol. 36, 734–747.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology; the need to count for prevalence. J. Appl. Ecol. 38, 921–931.

Martins, R., Stephan, N., Alves, R., 2008. Tubificidae (Annelida: Oligochaeta) as an indicator of water quality in an urban stream in Southeast Brazil. Acta Limnol. Bras. 20, 221–226.

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshw. Biol. 38, 237–246.

McCluskey, A., Lalkhen, A.G., 2007. Statistics II: central tendency and spread of data. Contin. Educ. Anaesthesia Crit. Care Pain 7, 127–130.

McCullagh, P.N.J.A., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London, UK.

McPherson, J., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography (Cop.) 30, 135–151.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.

Md Rawi, C., Al-Shami, S.A., Madrus, M.R., Ahmad, A.H., 2013. Biological and ecological diversity of aquatic macroinvertebrates in response to hydrological and physicochemical parameters in tropical forest streams of Gunung Tebu, Malaysia: implications for ecohydrological assessment. Ecohydrology 7, 496–507.

Özesmi, U., Mitsch, W.J., 1997. A spatial habitat model for the marsh-breeding red-winged blackbird (*Agelaius phoeniceus* L.) in coastal Lake Erie wetlands. Ecol. Model. 101, 139–152.

Park, Y.S., Cereghino, R., Compin, A., Lek, S., 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol. Model. 160, 265–280.

Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. Species: a spatial evaluation of climate impact on the envelope of species. Ecol. Model. 154, 289–300.

Poulsen, A.F., Viravong, S., 2001. Fish Migration and the Maintenance of Biodiversity in the Mekong River Basin.

Prendergast, J., Quinn, R., Lawton, J., Eversham, B., Gibbons, D., 1993. Rare species, the coincidence of diversity hotspots and cons. Nature 365, 335–336.

R Core Team, 2013. R: a Language and Environment for Statistical Computing.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33, 1689–1703.

Ripley, B., Venables, W., 2016. Package Nnet.

Rizo-Patrón, V.F., Kumar, A., McCoy Colton, M.B., Springer, M., Trama, F.A., 2013. Macroinvertebrate communities as bioindicators of water quality in conventional and organic irrigated rice fields in Guanacaste, Costa Rica. Ecol. Indic. 29, 68–78.

Roura-Pascual, N., Brotons, L., Peterson a, T., Thuiller, W., 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biol. Invasions 11, 1017–1031.

Schröder, W., Pesch, R., Schmidt, G., 2007. Statistical classification of terrestrial and marine ecosystems for environmental planning. Landsc. Online 2, 1–22.

Segurado, P., Araujo, M., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.

Sodhi, N.S., Koh, L.P., Brook, B.W., Ng, P.K.L., 2004. Southeast Asian biodiversity: an impending disaster. Trends Ecol. Evol. 19, 654–660.

Stockwell, D.R., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Model. 148, 1–13.

Stokland, J.N., Halvorsen, R., Stoa, B., 2011. Species distribution modelling—effect of design and sample size of pseudo-absence observations. Ecol. Model. 222, 1800–1809.

Suhaila, A., Che Salmah, M.R., 2014. Ecology of Ephemeroptera, Plecoptera and Trichoptera (Insecta) in rivers of the Gunung Jerai forest reserve: diversity and distribution of functional feeding groups. Trop. Life Sci. Res. 25, 61–73.

Therneau, T.M., Atkinson, E.J., 1997. An Introduction to Recursive Partitioning Using the Rpart Routines. Technical report no. 61, Rochester, Minnesota.

Toledo, M., Peña-Claros, M., Bongers, F., Alarcón, A., Balcázar, J., Chuviña, J., Leaño, C., Licona, J.C., Poorter, L., 2012. Distribution patterns of tropical woody species in response to climatic and edaphic gradients. J. Ecol. 100, 253–263.

Van Echelpoel, W., Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E., Mouton, A., Goethals, P.L.M., 2015. Species distribution models for sustainable ecosystem management. In: Park, Y., Lek, S., Baehr, C., Jorgensen, S. (Eds.), 19th Global Biennial Conference of the International-Society-for-Ecological-Modelling (ISEM). Elsevier Science BV, Toulouse France, pp. 115–134.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Zalinge, N., Van Thuok, N., 1998. It's big, unique and important: fisheries in the Lower Mekong Basin: as seen from a Cambodian perspective. Catch Cult. 4, 1–8.

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.

Article 4

Aquatic Invasions (2017) Volume 12, Issue 2: 147–158

DOI: https://doi.org/10.3391/ai.2017.12.2.03

Open Access

Research Article

Spatio-temporal co-occurrence of alien and native molluscs: a modelling approach using physical-chemical predictors

Ratha Sor[1,2,3,]*, Pieter Boets[2,4], Sovan Lek[1] and Peter L.M. Goethals[2]

[1]*Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS - Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cédex 4 – France*

[2]*Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure links 653, B9000 Ghent, Belgium*

[3]*Department of Biology, Faculty of Science, Royal University of Phnom Penh, Russian Boulevard, 12000, Phnom Penh, Cambodia*

[4]*Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000 Ghent, Belgium*

Author e-mails: sorsim.ratha@gmail.com (RS), pieter.boets@oost-vlaanderen.be (PB), sovannarath.lek@univ-tlse3.fr (SL), Peter.Goethals@UGent.be (PG)

**Corresponding author*

Received: 27 September 2017 / Accepted: 11 May 2017 / Published online: 23 May 2017

Handling editor: Kęstutis Arbačiauskas

Abstract

The invasion of alien species can have serious economic and ecological impacts. Ecologically, invasions often lead to an increased rate of native species replacement and decreased biodiversity. A critical step in the dominance of alien species is their successful co-occurrence with native species. In this study, we assessed the occurrence of alien molluscs and their co-occurrence with native molluscs and identified the determining physical-chemical variables. We expected that a combination of some key variables of water quality could provide suitable conditions promoting alien molluscs to occur and to co-occur with native molluscs. The analyses were based on 20-year data, collected from river systems across Flanders (Belgium). Classification Trees (CTs) were used to perform the analyses and to develop the predictive models. Based on CT models, the co-occurrence of alien and native molluscs could be reliably predicted based on physical-chemical variables. However, there was insufficient data to determine the environmental conditions in which alien taxa dominate. From the past to the present, spatial co-occurrence significantly increased. Sinuosity, ammonium and nitrate concentrations, chemical oxygen demand, pH and conductivity were the key determining variables. Our findings suggest that the co-occurrence of alien and native molluscs mainly occurs in straight rivers with good chemical water quality. These results provide insights into the ecology and behaviour of alien species which could support management practices and priority setting for conservation planning in surface waters of Flanders and Europe.

Key words: Invasion, habitat suitability, classification trees, species replacement, water quality, Flanders

Introduction

Invasive species have become a major concern for the global economy and environment (Sala et al. 2000). A large proportion of the economy has been spent on the management of agriculture, grassland and various natural ecosystems to mitigate the effects of alien invasive species (Williams et al. 2010; Hulme 2012). Moreover, their spread threatens native species of the same taxonomic groups and surrounding biotic communities via e.g. species replacement, food web reorganization and community composition simplification (Gurevitch and Padilla 2004; Bernauer and Jansen 2006; Didham et al. 2007). Once invasive species have successfully colonized new habitats and co-exist with native species, eradication is rarely possible (Regan et al. 2006). Consequently, the rate of replacement of native species by invasive species increases, which can thus lead to an overall decrease of native species (Olden et al. 2004).

Invasion success depends on traits of the invaders and the suitability of invaded environments (Kolar and Lodge 2001). Some taxa, e.g. *Corbicula* spp., are highly successful invaders due to their rapid spreading ability and their capability to withstand a wide range of environmental conditions (Werner and Rothhaupt 2007; Pigneur et al. 2014). Habitat modifications, resulting in changed physical and chemical conditions, often promote the local abundance and regional distribution of alien species (Didham et al. 2007). Increased trade (shipping) and improved chemical water quality may also promote the number and abundance of alien species (IKSR 2002; Boets et al. 2016). Therefore, identifying the environmental conditions in which alien species solely exist or co-exist with native species and determining those locations that could be invaded in the future will provide essential knowledge to support environmental management and conservation planning.

River systems in Europe have been exposed to the introduction of alien macroinvertebrate species. In the river Rhine, for example, alien species contribute 11.3% of the total macroinvertebrate species richness (Leuven et al. 2009). Among the macroinvertebrate invaders, molluscs constitute a large proportion (Leuven et al. 2009; Nunes et al. 2015). However, for most river systems in Europe, e.g. river systems in Flanders, the environmental conditions in which only alien species occur or the conditions preferred by both alien and native species (co-occurrence) are poorly studied. Recently, an inventory and habitat suitability model of alien macrocrustaceans in Flanders was conducted (Boets et al. 2013; Boets et al. 2016). Moreover, Boets et al. (2016) reported that alien mollusc species, e.g. the New Zealand mud snail (*Potamopyrgus antipodarum* J.E.Gray, 1843) and the acute bladder snail (*Physella acuta* (Draparnaud, 1805)), are highly abundant in the river systems of Flanders. As such, there is an urgent need to gain insight into the environmental conditions preferred by alien molluscs and to determine the conditions that allow alien molluscs to co-occur with native molluscs, as a basis for invasion control (e.g. locations and type of actions which deserve priority).

The aim of our study is to 1) provide an analysis of the spatio-temporal occurrence of alien molluscs and their co-occurrence with native molluscs in the river systems of Flanders over the past two decades (1991–2010), and 2) identify key determining physical-chemical variables associated with the sole occurrence of alien molluscs and their co-occurrence with native molluscs. We expected that a combination of some key variables of water quality could provide suitable conditions promoting alien molluscs to occur and to co-occur with native molluscs.

Material and methods

Data collection and treatment

The Flemish Environment Agency (VMM) has collected biological and environmental data in Flanders since 1989. The samples have been collected at more than 2500 sites spread over different water bodies. In this monitoring program, a standard handnet was used to collect macroinvertebrates following the method described by Gabriels et al. (2010). At sampling sites where the kick sampling method was not possible, artificial substrates were used. Seven alien and 27 native mollusc genera were identified. Electrical conductivity (EC), pH and dissolved oxygen (DO) were measured in the field with a hand-held probe (Cond 315i, oxi 330, wtw, Germany and 826 pH mobile, Metrohm, Switzerland). All additional chemical variables, i.e. ammonium (NH_4^+), chemical oxygen demand (COD), biological oxygen demand (BOD), total phosphorus (Pt), nitrate (NO_3^-), nitrite (NO_2^-), Kjeldahl nitrogen (KjN), and orthophosphate (oPO4), were retrieved from the monitoring dataset compiled by the VMM and which is accessible online (www.vmm.be). Nutrient analysis was performed spectrophotometrically in accordance with ISO 17025. GIS software (version 9.3.1) applied to the Flemish Hydrographic Atlas was used to determine the slope and the sinuosity of a watercourse using the difference in height in between two points 1000 m apart, and on a stretch of 100 m, respectively. For further detailed information on the determination of physical-chemical variables, we refer to Boets et al. (2016).

Data from 1991 to 2010 were used for the analyses. The data were divided into 4 periods, each encompassing 5 years of sampling effort (i.e. 1991 to 1995, 1996 to 2000, 2001 to 2005 and 2006 to 2010). This division provided more samples with which to model the preferred environmental conditions, and can provide useful information on changes in the occurrence of alien molluscs and their co-occurrence with native molluscs for each period. Due to limited frequency of occurrence for most alien mollusc genera (Table 1, 2), we decided to merge all alien genera to form one categorical variable. This provided us with a higher number of instances for our predictive models and thus a better and more robust development of the model. Moreover, we were not aiming to make predictions for individual taxa but rather to reveal common environmental conditions that most of the alien molluscs prefer. In the same way, all native genera were also merged to form one categorical variable. Environmental preferences of each genus of alien and

Table 1. The occurrence instances of each alien mollusc genus and of all genera that are merged together, compared to all collected samples (7695 samples) within the studied period (1991–2010). The sum of occurrences of different alien genera is 2522 instances, and the overlapping occurrences of alien genera are 494 instances.

Genus	*Corbicula*	*Dreissena*	*Ferrisia*	*Lithoglyphus*	*Menetus*	*Physella*	*Viviparus*	Merging all genera
Instances	130	745	381	30	4	1138	94	2028
Instances (%)	1.4	8.8	4.5	0.4	<0.1	13.9	1.1	26.4

Table 2. List of alien and native molluscs, their occurrences and abundance recorded for each period. The total number of samples for each period is indicated in brackets.

	Occurrence				Abundance			
	1991-1995	1996-2000	2001-2005	2006-2010	1991-1995	1996-2000	2001-2005	2006-2010
Taxa	(935)	(2021)	(2889)	(1850)	(935)	(2021)	(2889)	(1850)
Alien								
Corbicula	0	13	47	51	0	58	304	1092
Dreissena	59	132	257	227	683	1357	4980	29414
Ferrisia	10	17	199	120	125	40	1163	1133
Lithoglyphus	8	7	9	4	61	33	53	9
Menetus	0	0	0	4	0	0	0	11
Physella	2	1	236	831	13	2	15410	39361
Viviparus	15	27	32	12	142	162	148	126
Native								
Acroloxus	46	99	281	234	153	255	1760	1552
Ancylus	43	121	215	142	116	698	1097	1239
Anisus	152	318	472	235	633	2009	5493	2190
Anodonta	8	10	26	18	22	16	140	93
Aplexa	5	7	21	3	7	28	197	4
Armiger	25	63	260	148	63	150	1613	884
Bathyomphalus	84	189	237	159	447	915	1550	1864
Bithynia	225	498	808	514	2650	6639	27744	24220
Bythinella	3	4	0	0	33	16	0	0
Gyraulus	246	489	675	486	1598	2873	6471	11090
Hippeutis	27	50	206	131	121	131	1225	576
Lymnaea	512	1252	1677	918	6075	10458	23785	15381
Margaritifera	1	0	0	0	2	0	0	0
Marstoniopsis	1	0	0	0	1	0	0	0
Myxas	2	0	0	0	4	0	0	0
Physa	543	1103	1214	220	9358	13265	13918	4358
Pisidium	342	845	1272	799	3429	11237	36421	47265
Planorbarius	73	183	205	141	405	863	1297	948
Planorbis	113	230	280	195	786	1616	1841	2484
Potamopyrgus	165	306	529	365	3861	5424	28602	112402
Pseudamnicola	2	7	2	363	12	73	3	4088
Pseudanodonta	2	1	1	56	2	2	1	203
Segmentina	21	36	89	57	55	130	796	423
Sphaerium	224	451	873	526	2974	3386	7766	10878
Theodoxus	13	14	4	2	128	61	16	33
Unio	2	3	7	5	4	4	39	32
Valvata	242	445	734	458	3982	8193	25805	42883

native molluscs are provided in the Supplementary material Appendix 1. Each sampling site for each period was then categorized as: a "native" site (i.e. a site having only native molluscs present), an "alien" site (i.e. a site having only alien molluscs present) and a "co-occurrence" site (i.e. a site having both alien and native molluscs present), and this status was used as the response variable in analyses. Physical and chemical water quality variables that had missing values for more than 5% of the total samples were removed from the analyses. Each period thus consisted of one response categorical variable (native/alien/co-occurrence) and 13 predictor variables. To visualize the occurrence of alien molluscs and their co-occurrence with native molluscs, we produced an occurrence map using GIS-software (ArcGIS version 9.3.1). The summary of the physical and chemical water quality variables and of the response variable is shown in Table 3.

Table 3. Mean (and standard deviation) for environmental predictors and occurrence instances of each class of the response variable. NH_4^+: Ammonium, COD: Chemical Oxygen Demand, Pt: Total Phosphorus, EC: Electrical Conductivity, NO_3^-: Nitrate, NO_2^-: Nitrite, oPO_4: Orthophosphate, DO: Dissolved Oxygen. The number of sampled sites for each period is shown in square brackets.

Variable	Unit	Period 1991-1995 [509]	 1996-2000 [991]	 2001-2005 [1524]	 2006-2010 [1250]
NH_4^+	mg/L	2.6 (4.8)	1.9 (3.4)	1.6 (2.8)	2.3 (5.1)
COD	mg/L	55 (43)	39 (42)	34 (34)	36 (41)
Pt	mg/L	1.0 (1.4)	0.9 (1.4)	0.9 (2.3)	0.8 (1.0)
EC	µS/cm	1320 (2618)	987 (1149)	998 (1509)	921 (949)
NO_3^-	mg/L	3.3 (4.1)	4.0 (4.7)	3.6 (3.7)	3.0 (3)
NO_2^-	mg/L	0.2 (0.3)	0.2 (0.3)	0.2 (0.2)	0.2 (0.2)
oPO_4	mg/L	0.6 (0.9)	0.5 (1.1)	0.4 (0.7)	0.5 (0.8)
pH		7.5 (0.6)	7.6 (0.5)	7.7 (0.4)	7.6 (0.4)
DO	mg/L	7.5 (3.5)	6.8 (3)	6.9 (3.1)	6.6 (2.9)
Sinuosity		1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)
Slope	m/1000m	1.0 (1.4)	1.6 (2.8)	2.0 (3.6)	1.7 (2.8)
Response	class				
Co-occurrence		*66*	*146*	*567*	*875*
Alien		*6*	*15*	*63*	*133*
Native		*863*	*1860*	*2259*	*842*
Total instances		935	2021	2889	1850

Modelling

A Classification Tree (CT) model was used to predict the occurrence of native and alien molluscs and their co-occurrence, and to identify the determining physical-chemical variables. The CT was chosen among other machine learning approaches according to its performance for both predictive power and the importance of input variables (Chen et al. 2015). Moreover, this decision tree model is relatively simple to implement, easy to interpret, and it tolerates missing values during both the training and testing phases (Therneau and Atkinson 1997; De'ath and Fabricius 2000).

A CT model is based on growing and pruning to select an optimal tree. In the growing phase, the decision trees were fitted using a recursive partitioning algorithm. In each growing phase, the root of a tree (the initial node) is built from the most informative input variables. From the root, the data is split into left and right branches based on the splitting rules defined by the values of selected input variables. The growing process continues up to the terminal node until all the data in that node are of the same class or until some other stopping criterion is reached. The terminal nodes are called leaves and are labelled with the corresponding class (Quinlan 1986). In the pruning phase, the tree was pruned by setting a complex parameter at cp=0.05. Where the tree had only a root, we decreased the cp to a level (e.g. cp = 0.04, 0.03, 0.02, 0.01) that at least two terminal nodes were produced. In practice, the first few splits mostly provide a very informative division of the data (Therneau and Atkinson 1997). These criteria were set in order to make the trees easily interpretable and comparable in terms of the number of variables and complexity.

For each period, a three-fold cross-validation was used to train and validate the models. To build reliable models and to avoid misidentifying the key variables determining each class of the response variable, we made 3 replications of the three-fold cross validation. For each 3-fold cross-validation, the data was shuffled and randomly split into three subsets; two subsets were used for training and one subset for validation. For the second and third replication, we re-shuffled the data and randomly split it into new training and validation sets following the same procedures. From each training and validation set, a model was built and in this way, a performance value and the importance of each variable (as a percentage) of nine different models (3 models of each three-fold cross validation × 3 replications) were calculated. A mean performance value, obtained from the nine models, was used as a final criterion for model evaluation. Cohen's Kappa Statistic (K) and Correctly Classified Instances (CCI) were used to evaluate the model performance. The higher the value of K (ranging from 0 to 1) and of CCI (ranging from 0 to 100), the better the model predicts the response variable. The importance of each variable determining the preferred environmental conditions of each class (native/alien/co-occurrence) was averaged across the nine models. To identify which

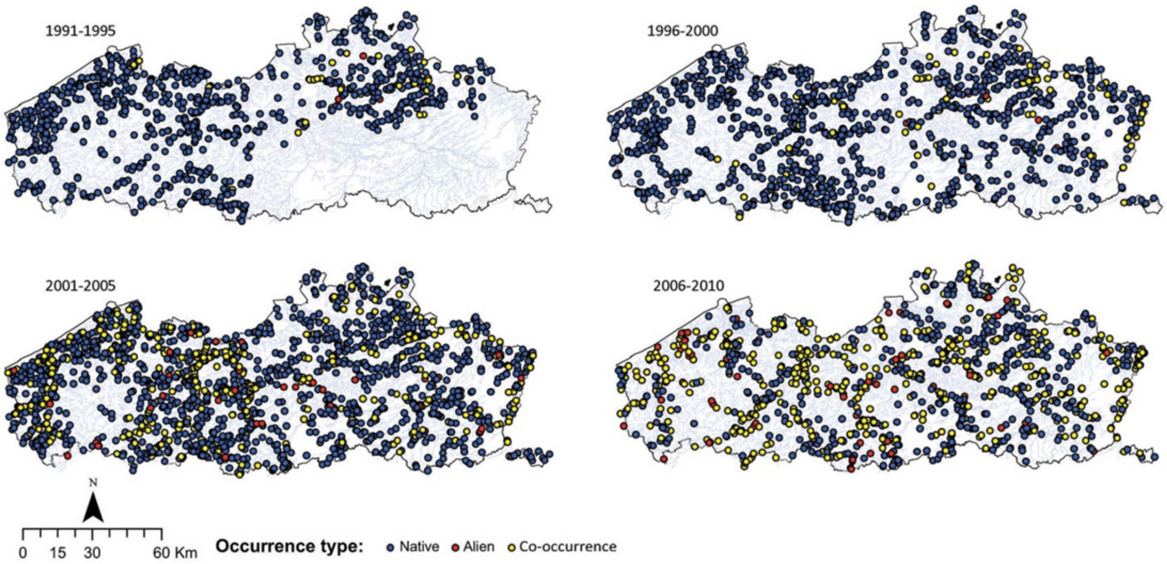


Figure 1. Sampling locations indicating the occurrence of native and alien molluscs and their co-occurrence for each period.

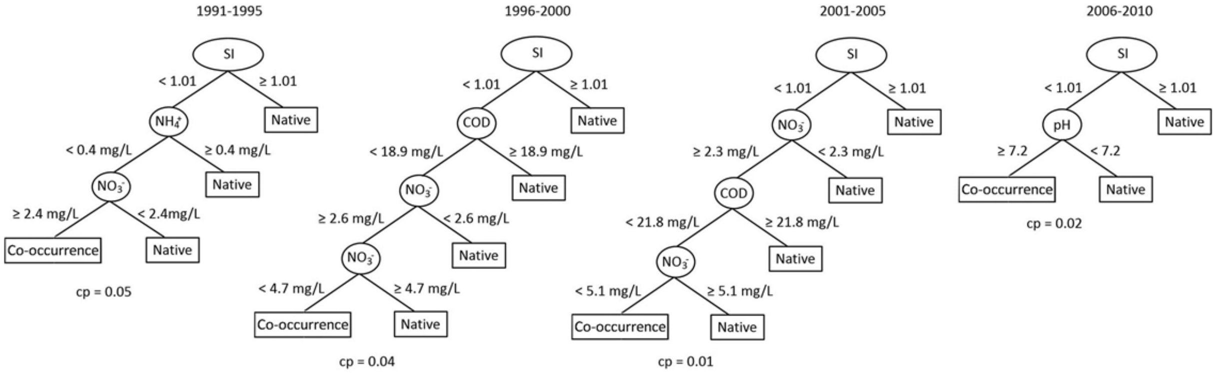


Figure 2. Classification trees predicting the "co-occurrence" and the "native" sites for each period. SI: Sinuosity, NH_4^+: Ammonium, NO_3^-: Nitrate, COD: Chemical Oxygen Demand, cp: pruning complex parameter.

variables significantly determine the preferred conditions of each class, the importance of each variable was compared based on the standard error. The same procedures and criteria were applied for modelling the data for each period across the whole studied period.

To develop and validate the models, and to construct the classification trees (for visualization), the package "rpart" in R (Breiman et al. 1984) was used. As we had many models which produced many trees for each period, we chose to construct the tree based on all data points of each period to be the representative one. All statistical analyses and calculations (K and CCI) were performed in R (R Core Team 2013).

Results

Occurrence of alien molluscs and their co-occurrence with natives

Overall, the occurrence of alien molluscs spatially and temporally increased in fluvial systems in Flanders (Figure 1). The "alien" sites accounted for 0.6%, 0.7%, 2.3% and 7.2% of the total samples for the period 1991–1995, 1996–2000, 2001–2005 and 2006–2010, respectively. The alien taxa which showed a notable increase in occurrences include *Corbicula, Dreissena, Ferrisia* and *Physella.* In the last period, a new alien genus (*Menetus*) was also recorded.

Detailed information on the occurrences and abundance of each alien mollusc genus is provided in Table 2. There was also a spectacular increase in the "co-occurrence" sites; they respectively accounted for 7.1% (66/935 samples), 7.2% (146/2020 samples), 21.1% (567/2689 samples) and 47.3% (875/1850 samples) for the corresponding periods (Table 3).

Key determining variables

For each period, the CT models were only able to reliably predict the "co-occurrence" and the "native" sites. The representative trees of the CT models for each period are shown in Figure 2. Sinuosity was always one of the most important variables for the models of each period and this, together with chemical water quality variables (e.g. NH_4^+, NO_3^-, COD, pH, Figure 2) and EC, were the key factors determining the predictive models (Figure 2, 3). When sinuosity was lower than 1.01, co-occurrence between alien and native molluscs was evident where NH_4^+ <0.4 mg/L and NO_3^- ≥2.4 mg/L for the period 1991–1995; COD <18.9 mg/L and NO_3^- in between 2.6–4.7 mg/L for 1996–2000; COD <21.8 mg/L and NO_3^- in between 2.3–5.1 mg/L for 2001–2005; and where pH ≥7.2 for 2006–2010 (Figure 2). Sites where only alien molluscs occurred could not be reliably predicted based on the available physical-chemical data (Figure 4).

Model performance

The mean Cohen's Kappa Statistic (K) and the mean overall Correctly Classified Instances (Overall-CCI) of the models decreased from the past to the most recent period (Figure 5A–B). For the four periods (1991–1995, 1996–2000, 2001–2005 and 2006–2010), the corresponding mean K was 0.34, 0.32, 0.22 and 0.16, and the corresponding mean Overall-CCI was 93%, 92%, 79% and 54%. The mean CCI of models predicting the "co-occurrence" sites (CCI-co-occurrence) was lower for the first 3 periods (28%, 25% and 23%), while it was higher for the most recent period (52%, Figure 5C). On the contrary, the mean CCI of models predicting the "native" sites (CCI-Native) for the past 3 periods (99%, 98%, and 96%) was substantially higher compared to that of the most recent period (66%, Figure 5D). The models predicting the "alien" sites for the 4 periods did not yield any reliable prediction. Only one model that was based on the data from the latest period correctly predicted one instance of "alien" sites. The confusion matrices obtained from the models and which were used to calculate the model performance measures (K and CCI) are provided in Appendix 2.

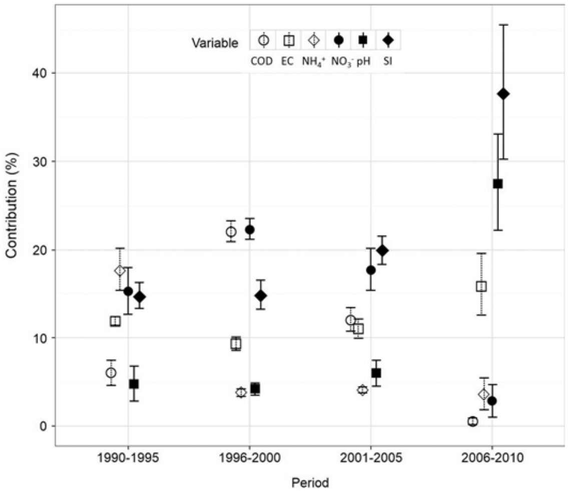


Figure 3. Mean and standard error bar showing the importance of the variables contributing most to the predictive models for each period. COD: Chemical Oxygen Demand, EC: Electrical Conductivity, NH_4^+: Ammonium, NO_3^-: Nitrate, SI: Sinuosity. Variables that contributed less to the predictive models are not shown.

Discussion

Occurrence of alien molluscs and their co-occurrence with the natives

Alien molluscs have spread spectacularly in several European river systems during the last few decades, e.g. the rivers Rhine and Meuse (Bernauer and Jansen 2006; Collas et al. 2014; Pigneur et al. 2014). This phenomenon is similarly observed in our study in the rivers in Flanders. The remarkable increase in the occurrence of alien molluscs over the past two decades may reveal their outbreak or invasion success across the fluvial systems in Flanders, which resulted in an increased incidence of co-occurrence of alien and native molluscs.

Key determining variables

CT models indicated that co-occurrence is mainly determined by sinuosity and by a set of chemical water quality variables (i.e. NH_4^+, NO_3^-, COD, pH and EC). Sinuosity was always one of the most important factors determining co-occurrence as it formed the main root in all models. Sites having a low sinuosity (<1.01), which corresponds to mainly straight rivers, may be subjected to a high number of passing ships, which is considered one of the main pathways of invasions (Bij de Vaate et al. 2002; Nunes et al. 2015). Straight rivers shorten travelling distances, resulting in more frequent transportation,

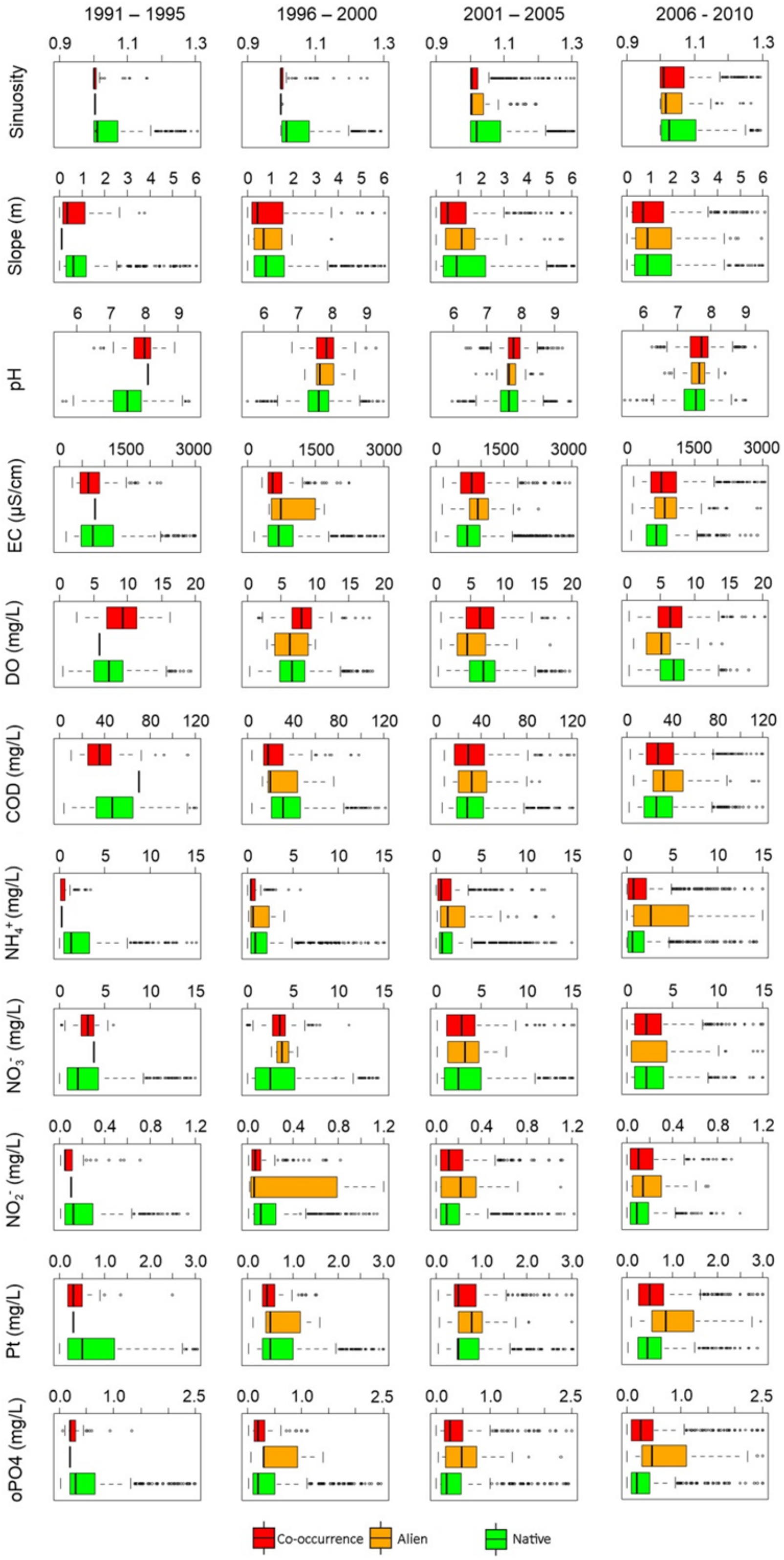


Figure 4. Box and whisker plots of physical-chemical variables in which each occurrence type occurred. Rectangles show first and third quartiles, dark bars are the medians, the lower and upper bars are the minimum and maximum values, and the circles are outliers. EC: Electrical Conductivity, DO: Dissolved Oxygen, COD: Chemical Oxygen Demand, NH_4^+: Ammonium, NO_3^-: Nitrate, NO_2^-: Nitrite, Pt: Total Phosphorus, oPO4: Orthophosphate.

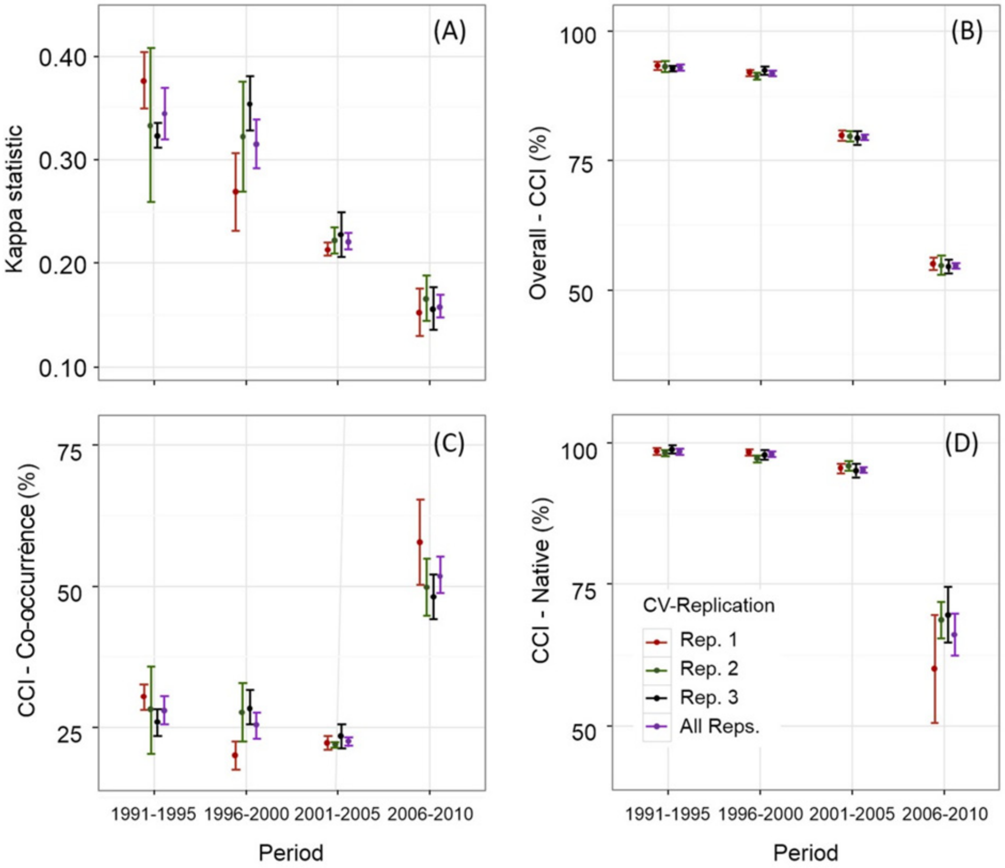


Figure 5. Mean and standard error bar indicating the overall model performance based on Cohen's Kappa statistic (A) and Correctly Classified Instances (CCI: B), and the predictive power of the predicted class "co-occurrence" (C) and the class "native" (D). CV: cross-validation, Reps: Replications.

thus allowing a large and frequent amount of ballast water to be released. Consequently, with a higher number of introductions, the survival rate of alien molluscs increases (Gollasch 2006). We found that the hotspots of mollusc invasion (the alien and the co-occurrence sites) were mainly situated in brackish polder watercourses and in large rivers that have a shorter distance to the ports in the Rhine delta, to the coast of the North Sea and to other large rivers (e.g. Meuse River). This observation is supported by evidence in Boets et al. (2016) and in Grabowski et al. (2009) who found that alien fauna mostly inhabited large rivers where intensive navigation takes place.

Besides river morphology, water quality is one of the major factors influencing the distribution and diversity of freshwater fauna (Leuven et al. 2009; Wang et al. 2012). Key variables used to evaluate water quality are NH_4^+, NO_3^+, COD, pH, DO, Chloride, and total phosphorus (US-EPA 1986; SEQ-Eau 2003; WWF 2007; Chea et al. 2016). Good water quality supports a high diversity of invertebrates (Leuven et al. 2009). In our study, where sampling sites had a low sinuosity (<1.01), co-occurrence was mainly dependent on chemical water quality status.

River systems in Europe as well as in Flanders have suffered from severe water quality degradation in previous decades. During these periods, some native species were reported to disappear (Bernauer and Jansen 2006) and only species that were able to withstand this water quality degradation remained. In the early 1990s, the invasion of alien molluscs seemed to be at an initial stage as reflected by the presence of only five alien taxa with limited spatial occurrence and low abundance (Figure 1, Table 2). Although water quality was degraded during these periods, we found that a low NH_4^+ concentration (<0.4 mg/L) and a nitrate concentration higher than 2.4 mg/L were preferred by both the alien and remaining native molluscs. This could be because higher NH_4^+ concentrations negatively affect their living conditions (Friberg et al. 2010). Moreover, an enrichment of nutrient content (e.g. NO_3^- >2.4 mg/L) could be advantageous for alien molluscs at the beginning of the invasion. Nutrient enrichment provides resources that can be used by alien molluscs and thus may enhance their proliferation (Hall et al. 2003; Strayer 2010). However, a very high nutrient concentration (e.g. NO_3^- >25 mg/L) is also an indication of a high level of water pollution (SEQ-Eau 2003), which can negatively affect both alien and native species (Boets et al. 2013). Nonetheless, a low NH_4^+ concentration and an optimum nutrient content (i.e. mean NO_3^- of 3.3 mg/L) facilitated alien and native molluscs co-existence during the early 1990s.

The late 1990s and early 2000s appeared to be an expansion phase for some alien molluscs (e.g. *Ferrisia, Physella, Dreissena* and *Corbicula*) since we found a large increase in their occurrences and abundance (Table 2). This increase may result in a substantial effect on water quality in river systems. A high abundance of filter-feeders (e.g. *Dreissena* and *Corbicula*) and *Physella,* which also feeds on phytoplankton, can lead to increased nutrient concentration. This is because the filter feeders consume phytoplankton, thus limiting the abundance of nutrient utilizing phytoplankton communities, allowing nutrient inputs (e.g. nitrate and phosphate) from surrounding areas to increase in river systems (Lavrentyev et al. 2000; Pigneur et al. 2014). However, this might not have led to a severe impact on water quality because during these periods there was a successful rehabilitation program for improving water quality, restoring riverine ecosystems and improving habitat connectivity in Europe (Leuven et al. 2009). A decreasing trend of nitrate and other water quality variables (e.g. COD) was observed in the river systems in Flanders (UN 2004). Therefore, it can be inferred that improved water quality and the rehabilitation programs not only helped to recover the diversity of native species but also promoted the occurrence and abundance of alien species. Indeed, Leuven et al. (2009) indicated that when hydromorphological conditions remain unchanged, improvement in water quality promotes alien species.

In the late 2000s, a high percentage of surface water bodies (43%) in Europe were considered to be at a “good status”. The number of waterbodies increased to 53% in 2015, and Flanders was one of the regions that well implemented the policy of the Water Framework Directive (EU 2015). The improvement in water quality was associated with a great increase in the spatial occurrence and abundance of alien molluscs. The two filter feeders (i.e. *Corbicula* and *Dreissena*) always expanded their range and frequency in the late 2000s even though the number of sampling sites and total sample size were lower than the early 2000s. Moreover, a new alien genus, *Menetus,* also emerged in the late 2000s. This suggests that the late 2000s can be considered as the expansion phase of the existing alien molluscs and the beginning of the expansion phase of the recently introduced alien genus, *Menetus.* These results are therefore unlikely to be an effect of the sampling strategy, but rather reflect the suitable physical-chemical conditions for alien molluscs to spread and proliferate. However, we found that the pH value, which was the second most important variable in determining the co-occurrence of alien and native molluscs (Figure 2, 3) during this period, was relatively high (pH: 6.5–8.5). This is probably related to certain specific environmental conditions linked to geographic regions. For example, a high pH value could be mainly recorded from brackish polder watercourses and from the main harbour watercourses where a high level of seawater intrusion occurs and a high intensity of human-related activities takes place. The alkaline watercourses may have a higher concentration of calcium and magnesium compared to inland watercourses, thus may be preferred by some alien molluscs (e.g. *Ferrisia* and *Physella)* since they require a large amount of calcium and magnesium to form their shells (Brodersen and Madsen 2003; Gallardo and Aldridge 2013). Moreover, molluscs in the family Physidae, including the alien *Physella* and the native *Physa*, are pollution tolerant and can occur in areas with high pH values (Rodrigues Capítulo et al. 2001; De Troyer et al. 2016). Boets et al. (2013) also found an increase of alien species abundance with increasing pH.

Across the models of the four periods, other than NH_4^+, NO_3^-, COD and pH, EC was also selected as one of the key determining variables. Many studies have shown that high conductivity mostly favours alien species. For example, alien amphipods in the Vistula and Oder rivers of Poland, alien macrocrustaceans in fluvial systems in Flanders and alien gastropods in isolated ponds in Poland all benefited from high conductivity (Grabowski et al. 2009; Boets et al. 2013; Gallardo and Aldridge 2013; Spyra and Strzelec 2014). It is likely that alien mollusc species are well adapted to withstand high values of EC, while only a few native species (e.g. *Lymnaea stagnalis* (Linnaeus, 1758)), which occur in a wide range of environmental conditions (Brown et al. 2011), are able to co-exist in these areas. Our findings suggest that increased spatial co-occurrence results from the introduction or migration of alien species to the connected environments where native species are present or to new environments where conditions fit them best, e.g. sites having a high conductivity. Colonizing new niches where most of the native species are not able to thrive is one of the main strategies found among invaders (Verbrugge et al. 2012). This might further imply that, after success in co-existing with the native species, the aliens may overrun some native species and further spread to new areas. Clear evidence can be seen from the drastically increased occurrences and abundance of most alien molluscs from the past to the present, from the decrease in occurrences and abundance of the native *Physa, Aplexa* and *Theodoxus*, and from the disappearance of the native *Bythinella, Margaritifera, Marstoniopsis* and *Myxas* (Table 2).

Although not directly taken into account in our models, previous studies have demonstrated that dispersal vectors are important in making predictions on future locations that may be invaded by alien species. Indeed, recent research on the dispersal of alien macrocrustaceans in Flanders (Boets et al. 2013) showed that increased shipping and the connection between waterways promote the dispersal of alien species. Moreover, habitat conditions (bank structures and substrates) and hydrological variables (e.g. distance to ports/coast, flow regime and connectivity) can also influence the occurrence of alien species (Josens et al. 2005; Messiaen et al. 2010). However, further research suggested that although habitat and hydrological variables can improve model reliability when predicting the spreading rate of alien species, these variables are often not the limiting factor when making predictions on the scale of Flanders (Boets et al. 2014). Nevertheless, these variables should be taken into account for future research that aims to analyse and predict particular preferred conditions of alien molluscs at a larger scale.

Model performance

Predictive models are widely applied to assess the environments or areas that alien species have invaded or would invade (Pitt et al. 2009; Boets et al. 2013; Chen et al. 2015). In many cases, the performance of these predictive models ranges from fair to moderate (Gabriels et al. 2007; Boets et al. 2013). In our study, the overall performance of the models was moderate to good. When using data from the past period, the performance was higher than when using data from the most recent period (Figure 5A–B). This could be attributed to less complex biotic interactions and to the limited frequency of occurrences of alien species in the past. In this context, alien species may invade those sites where competition is low and where few native species occur. As alien species start to spread, competition with native species increases, and thus some native species which had a small environmental range may disappear (i.e. *Bythinella, Margaritifera, Marstoniopsis, Myxas*) or decrease their occurrence and abundance (e.g. *Aplexa, Theodoxus*). This is epitomized by alien and native molluscs in the family Physidae. In the past, the alien *Physella* occurred at only few sites with a few individuals, while the native *Physa* abundantly and widely occurred. A contrasting relationship between the two taxa was observed for the last period, in part due to a declining trend in the abundance and spatial occurrence of the native *Physa*. This could be a result of competition for food and niche with the alien *Physella*. Biotic interactions are important in predicting species distribution (Araújo and Luoto 2007; Meier et al. 2010), and when included will generally increase the predictive performance of a model. The exclusion of biotic interactions in our study may explain the overall lower performance of the models based on the most recent period. Moreover, during the late 2000s the range of environmental conditions where alien species occur increased (Figure 4) while many native species recovered their range and density (Table 2), due to improved water quality. This higher co-occurrence might also explain the lower performance of the models.

Similarly, higher co-occurrence may have influenced the higher reliability in the prediction of the "co-occurrence" of alien and native molluscs for the recent period compared to the first three periods. Likewise, models predicting the "native" sites yielded a high performance when the prediction was based on the periods (i.e. the first three periods) that have a large sample size of the "native" sites. This is quite logical as for predictive models the more input samples provided the better the models learn, and as a result, a higher predictive performance can be obtained (Stockwell and Peterson 2002; Hernandez et al. 2006). On the other hand, the models were not able to predict the "alien" sites (Appendix 2), due to a low number of instances of this particular class. Although the number of samples of the "alien" sites increased in the most recent period, it was still not sufficient for the models to learn and make a correct prediction. Small sample sizes, together with the opportunistic and generalistic characteristics of the alien species (Nehring 2006), are therefore considered the main reasons for the models to yield a very low performance. Additional observations or a particular optimization approach is thus recommended to better predict the "alien" sites and evaluate the predictive power of the models.

Conclusion

From the past to the most recent situation, there is an increasing trend in the spatial co-occurrence of alien and native molluscs in Flanders. Co-occurrence was predicted to mainly occur in rivers having low sinuosity and good chemical water quality. In addition, our most recent data indicated that alien molluscs have reached a relatively high number of sites where natives were not present, indicating either that alien molluscs have invaded more new sites or replaced native species at sites where they previously occurred. Given that our models were not able to make reliable predictions for environmental conditions preferred by alien molluscs, additional predictors and observations or perhaps a particular optimization approach is needed to predict the habitat conditions where alien molluscs are able to dominate the community. These results provide important information regarding the past and current co-existence of alien and native molluscs in Flanders. Our findings may be used to support management and conservation planning.

Acknowledgements

We are grateful to the Flemish Environment Agency (VMM) for the opportunity to study their samples. The first author would like to sincerely thank Erasmus Mundus LOTUS project (European Commission) and the Forum Belmont TLSCC project for providing the possibility to do this research. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Araújo MB, Luoto M (2007) The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* 16: 743–753, https://doi.org/10.1111/j.1466-8238.2007.00359.x

Bernauer D, Jansen W (2006) Recent invasions of alien macro-invertebrates and loss of native species in the upper Rhine River, Germany. *Aquatic Invasions* 1: 55–71, https://doi.org/10.3391/ai.2006.1.2.2

Bij de Vaate A, Jazdzewski K, Ketelaars H, Gollasch S, Van der Velde G (2002) Geographical patterns in range extension of Ponto-Caspian macroinvertebrate species in Europe. Canadian *Journal of Fisheries and Aquatic Sciences* 59: 1159–1174, https://doi.org/10.1139/f02-098

Boets P, Lock K, Goethals PLM (2013) Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. *Ecological Informatics* 17: 73–81, https://doi.org/10.1016/j.ecoinf.2012.06.001

Boets P, Pauwels IS, Lock K, Goethals PLM (2014) Using an integrated modelling approach for risk assessment of the 'killer shrimp' *Dikerogammarus villosus*. *River Research and Applications* 30: 403–412, https://doi.org/10.1002/rra.2658

Boets P, Brosens D, Lock K, Adriaens T, Aelterman B, Mertens J, Goethals PLM (2016) Alien macroinvertebrates in Flanders (Belgium). *Aquatic Invasions* 11: 131–144, https://doi.org/10.3391/ai.2016.11.2.03

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth International Group, Betmont, California, 368 pp

Brodersen J, Madsen H (2003) The effect of calcium concentration on the crushing resistance, weight and size of *Biomphalaria sudanica* (Gastropoda: Planorbidae). *Hydrobiologia* 490: 181–186, https://doi.org/10.1023/A:1023495326473

Brown R, Soldánová M, Barrett J, Kostadinova A (2011) Small-scale to large-scale and back: larval trematodes in *Lymnaea stagnalis* and *Planorbarius corneus* in Central Europe. *Parasitology Research* 108: 137–150, https://doi.org/10.1007/s00436-010-2047-z

Chea R, Grenouillet G, Lek S (2016) Evidence of water quality degradation in Lower Mekong Basin revealed by Self Organizing Map. *PLoS ONE* 11: e0145527, https://doi.org/10.1371/journal.pone.0145527

Chen L, Peng S, Yang B (2015) Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. *Biological Invasions* 17: 2187–2198, https://doi.org/10.1007/s10530-015-0870-y

Collas FPL, Koopman KR, Hendriks AJ, van der Velde G, Verbrugge LNH, Leuven RSEW (2014) Effects of desiccation on native and non-native molluscs in rivers. *Freshwater Biology* 59: 41–55, https://doi.org/10.1111/fwb.12244

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–3192, https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

De Troyer N, Mereta S, Goethals P, Boets P (2016) Water quality assessment of streams and wetlands in a fast growing East African City. *Water* 8: 123, https://doi.org/10.3390/w8040123

Didham RK, Tylianakis JM, Gemmell NJ, Rand T, Ewers RM (2007) Interactive effects of habitat modification and species invasion on native species decline. *Trends in Ecology and Evolution* 22: 489–96, https://doi.org/10.1016/j.tree.2007.07.001

EU (2015) 4th European Water Conference Conference report. Brussels, 45 pp

Friberg N, Skriver J, Larsen SE, Pedersen ML, Buffagni A (2010) Stream macroinvertebrate occurrence along gradients in organic pollution and eutrophication. *Freshwater Biology* 55: 1405–1419, https://doi.org/10.1111/j.1365-2427.2008.02164.x

Gabriels W, Goethals PLM, Dedecker AP, Lek S, De Pauw N (2007) Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. *Aquatic Ecology* 41: 427–441, https://doi.org/10.1007/s10452-007-9081-7

Gabriels W, Lock K, De Pauw N, Goethals PLM (2010) Multimetric Macroinvertebrate Index Flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium). *Limnologica* 40: 199–207, https://doi.org/10.1016/j.limno.2009.10.001

Gallardo B, Aldridge DC (2013) Priority setting for invasive species management: risk assessment of Ponto-Caspian invasive species into Great Britain. *Ecological Applications* 23: 352–364, https://doi.org/10.1890/12-1018.1

Gollasch S (2006) Overview on introduced aquatic species in European navigational and adjacent waters. *Helgoland Marine Research* 60: 84–89, https://doi.org/10.1007/s10152-006-0022-y

Grabowski M, Bacela K, Konopacka A, Jazdzewski K (2009) Salinity-related distribution of alien amphipods in rivers provides refugia for native species. *Biological Invasions* 11: 2107–2117, https://doi.org/10.1007/s10530-009-9502-8

Gurevitch J, Padilla DK (2004) Are invasive species a major cause of extinctions? *Trends in Ecology and Evolution* 19: 470–474, https://doi.org/10.1016/j.tree.2004.07.005

Hall RO, Tank JL, Dybdahl MF (2003) Exotic snails dominate nitrogen and carbon cycling in a highly productive system. *Frontiers in Ecology and the Environment* 1: 407–411, https://doi.org/10.1890/1540-9295(2003)001[0407:ESDNAC]2.0.CO;2

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29: 773–785, https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hulme PE (2012) Weed risk assessment: a way forward or a waste of time? *Journal of Applied Ecology* 49: 10–19, https://doi.org/10.1111/j.1365-2664.2011.02069.x

IKSR (2002) Das Makrozoobenthos des Rheins 2000. Koblenz, 46 pp

Josens G, De Vaate AB, Usseglio-Polatera P, Cammaerts R, Chérot F, Grisez F, Verboonen P, Bossche JPV (2005) Native and exotic Amphipoda and other Peracarida in the River Meuse: new assemblages emerge from a fast changing fauna. *Hydrobiologia* 542: 203–220, https://doi.org/10.1007/s10750-004-8930-9

Kolar CS, Lodge DM (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology and Evolution* 16: 199–204, https://doi.org/10.1016/S0169-5347(01)02101-2

Lavrentyev PJ, Gardner WS, Yang LY (2000) Effects of the zebra mussel on nitrogen dynamics and the microbial community at the sediment-water interface. *Aquatic Microbial Ecology* 21: 187–194, https://doi.org/10.3354/ame021187

Leuven RSEW, van der Velde G, Baijens I, Snijders J, van der Zwart C, Lenders HJR, bij de Vaate A (2009) The river Rhine: a global highway for dispersal of aquatic invasive species. *Biological Invasions* 11: 1989–2008, https://doi.org/10.1007/s10530-009-9491-7

Meier ES, Kienast F, Pearman PB, Svenning JC, Thuiller W, Araújo MB, Guisan A, Zimmermann NE (2010) Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography* 33: 1038–1048, https://doi.org/10.1111/j.1600-0587.2010.06229.x

Messiaen M, Lock K, Gabriels W, Vercauteren T, Wouters K, Boets P, Goethals PLM (2010) Alien macrocrustaceans in freshwater ecosystems in the eastern part of Flanders (Belgium). *Belgian Journal of Zoology* 140: 30–39

Nehring S (2006) Four arguments why so many alien species settle into estuaries, with special reference to the German river Elbe. *Helgoland Marine Research* 60: 127–134, https://doi.org/10.1007/s10152-006-0031-x

Nunes AL, Tricarico E, Panov VE, Cardoso AC, Katsanevakis S (2015) Pathways and gateways of freshwater invasions in Europe. *Aquatic Invasions* 10: 359–370, https://doi.org/10.3391/ai.2015.10.4.01

Olden JD, Poff NL, Douglas MR, Douglas ME, Fausch KD (2004) Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology and Evolution* 19: 18–24, https://doi.org/10.1016/j.tree.2003.09.010

Pigneur L-M, Falisse E, Roland K, Everbecq E, Deliège JF, Smitz, JS, Van Doninck K, Descy JP (2014) Impact of invasive Asian clams, *Corbicula* spp., on a large river ecosystem. *Freshwater Biology* 59: 573–583, https://doi.org/10.1111/fwb.12286

Pitt J, Worner S, Suarez AV (2009) Predicting Argentine ant spread over the heterogeneous landscape using a spatially explicit stochastic model. *Ecological Applications* 19: 1176–1186, https://doi.org/10.1890/08-1777.1

Quinlan JR (1986) Induction of Decision Trees. *Machine Learning* 1: 81–106, https://doi.org/10.1007/BF00116251

R Core Team (2013) R: a language and environment for statistical computing. http://www.r-project.org

Regan TJ, McCarthy MA, Baxter PWJ, Dane Panetta F, Possingham HP (2006) Optimal eradication: when to stop looking for an invasive plant. *Ecology Letters* 9: 759–766, https://doi.org/10.1111/j.1461-0248.2006.00920.x

Rodrigues Capítulo A, Tangorra M, Ocón C (2001) Use of benthic macroinvertebrates to assess the biological status of Pampean streams in Argentina. *Aquatic Ecology* 35: 109–119, https://doi.org/10.1023/A:1011456916792

Sala OE, Chapin III FS, Armesto JJ, Berlow E, Bloomfield J, Dirzo R, Huber-Sanwald E, Huenneke LF, Jackson RB, Kinzig A, Leemans R, Lodge DM, Mooney HA, Oesterheld M, Poff NL, Skykes MT, Walker BH, Walker M, Wall DH (2000) Global biodiversity scenarios for the year 2100. *Science* 287: 1770–1774, https://doi.org/10.1126/science.287.5459.1770

SEQ-Eau (2003) Système d'évaluation de la qualité de l'eau des cours d'eau. Agences de l'eau-Minist ère de l'écologie et du développement durable, 40 pp

Spyra A, Strzelec M (2014) Identifying factors linked to the occurrence of alien gastropods in isolated woodland water bodies. *Naturwissenschaften* 101: 229–239, https://doi.org/10.1007/s00114-014-1153-7

Stockwell DR, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1–13, https://doi.org/10.1016/S0304-3800(01)00388-X

Strayer DL (2010) Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. *Freshwater Biology* 55: 152–174, https://doi.org/10.1111/j.1365-2427.2009.02380.x

Therneau TM, Atkinson EJ (1997) An introduction to recursive partitioning using the rpart routines. Technical report no. 61. Rochester, Minnesota, 67 pp

UN (2004) Freshwater Country Profile-Belgium, 29 pp, http://www.google.be/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwj2haLAmqDOAhVHGsAKHSqrAHUQFggcMAA&url=http://www.un.org/esa/agenda21/natlinfo/countr/belgium/belgiumwater04f.pdf&usg=AFQjCNFyyRBxfJJxe5tFoXKJJjLM-Xiq4w&bvm=bv.128617741,d.ZGg (accessed on 1 August 2016)

US-EPA (1986) Quality criteria for water. Office of Water Regulations and Standards. Washington D.C., 450 pp

Verbrugge LNH, Schipper AM, Huijbregts MAJ, van der Velde G, Leuven RSEW (2012) Sensitivity of native and non-native mollusc species to changing river water temperature and salinity. *Biological Invasions* 14: 1187–1199, https://doi.org/10.1007/s10530-011-0148-y

Wang B, Liu D, Liu S, Zhang Y, Lu D, Wang L (2012) Impacts of urbanization on stream habitats and macroinvertebrate communities in the tributaries of Qiangtang River, China. *Hydrobiologia* 680: 39–51, https://doi.org/10.1007/s10750-011-0899-6

Werner S, Rothhaupt KO (2007) Effects of the invasive bivalve *Corbicula fluminea* on settling juveniles and other benthic taxa. *Journal of the North American Benthological Society* 26: 673–680, https://doi.org/10.1899/07-017R.1

Williams F, Eschen R, Harris A, Djeddour D, Pratt C, Shaw R, Varia S, Lamontagne-Godwin J, Thomas S, Murphy S (2010) The economic cost of invasive non-native species on Great Britain. Wallingford, 199 pp

WWF (2007) National surface water classification criteria & irrigation water quality guidelines for Pakistan. Lahore, Pakistan. WWF-Pakistan, 33 pp

Supplementary material

The following supplementary material is available for this article:

Appendix 1. Mean as well as minimum and maximum values of each physical-chemical variable measured at sites where the presence of each genus of alien and native molluscs was recorded.

Appendix 2. Confusion matrices showing the observed and predicted classes obtained from Classification Tree models of the four periods.

This material is available as part of online article from:

http://www.aquaticinvasions.net/2017/Supplements/AI_2017_Sor_etal_Supplement.pdf

Title: Optimizing the reliability of classification tree models in predicting alien mollusc occurrence: a hindcasting- and forecasting-based approach

Authors:
Ratha Sor[1,2,3,*], correspondence, e-mail: sorsim.ratha@gmail.com
Pieter Boets[2,4], e-mail: pieter.boets@oost-vlaanderen.be
Sovan Lek[1], e-mail: sovannarath.lek@univ-tlse3.fr
Peter Goethals[2], e-mail: Peter.Goethals@UGent.be

Institutional address:
[1] Université de Toulouse, Laboratoire Evolution & Diversité Biologique, UMR 5174, CNRS - Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cédex 4 – France.
[2] Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Campus Coupure building F, Coupure links 653, B9000 Ghent, Belgium.
[3] Department of Biology, Faculty of Science, Royal University of Phnom Penh, Russian Boulevard, 12000, Phnom Penh, Cambodia.
[4] Provincial Centre of Environmental Research, Godshuizenlaan 95, 9000 Ghent, Belgium.

Abstract

Using reliable models to potentially forecast alien species occurrence is a pragmatic approach that is often used for assessment and management purposes. Models' reliability can be tested via hindcasting – a predicting of what happened in the past. Using 7695 field data points, collected over two decades, and the development of cloned data, we optimized the reliability of a classification tree (CT) model in hind- and forecasting alien mollusc occurrence in river systems in Flanders (Belgium). Random and stratified resampling approaches were used to split the data, which was used to develop calibrated and validated CT models with different parameterizations. We showed that the CT models, which were calibrated and validated using only field data are unable to predict the past and present putative occurrence of alien molluscs. We then optimized the CT models based on a combined dataset, being referred to the combination of field data and cloned data (i.e. a dataset obtained by independently duplicating the filed data points by *k* different individuals). This optimization demonstrated the capability the CT models in predicting alien mollusc occurrence with a low error rate. This result corresponds to field observations, where alien mollusc occurrence has been observed over the last two decades in Flanders. With the same parameterization, the two resampling approaches (i.e. random vs. stratified) used in the calibration and validation process are unlikely to affect the reliability of CT models. Our finding reports the robustness of the CT models in predicting alien species occurrence, making them suitable to be applied for assessment and management of alien species.

Keywords: Classification trees, predictive models, water quality, resampling approach, cloned data.

INTRODUCTION

Freshwater ecosystems are exposed to a high degree of isolation and endemism, and to a range of environmental and human pressures, which make these systems vulnerable to biological invasions (Dudgeon et al., 2006; Richter et al., 1997). The distribution of alien invasive species in freshwater ecosystems is increasing and is becoming one of the leading concerns for global economy and biodiversity (Sala et al., 2000). Major ecological effects of invasions cover replacement of native species, changes in nutrient cycling, reorganization of food webs and simplification of community composition (Bernauer and Jansen, 2006; Didham et al., 2007; Gurevitch and Padilla, 2004), all of which in turn alter the natural mechanisms of ecosystem processes.

Insights for practical assessment and management of invasive species have been revealed from previous studies which are based on predictive models (Boets et al., 2013; Chen et al., 2015; Elith and Leathwick, 2009; Hill et al., 2013). For example, modelling approaches have been used to predict distributional areas of invasive/alien plants, herbs and macroinvertebrates (Boets et al., 2013; Chen et al., 2015; Roura-Pascual et al., 2009; Thuiller et al., 2005; Vicente et al., 2011). Approaches employed include generalized linear models, generalized additive models, classification and regression trees, random forest and artificial neural networks, all of which are data-driven or knowledge based, and are considered important tools for predicting species distributions (Guisan and Thuiller, 2005).

While many available predictive approaches are widely implemented, a wide variation of their performance has been frequently reported (Elith et al., 2006; Guisan et al., 2007; Segurado and Araujo, 2004). Some models even yield contrasting predictions of habitat suitability (e.g. Guisan *et al.* 2007; Evangelista *et al.* 2008; Roura-Pascual *et al.* 2009). As such, when calibrating and validating predictive models, data characteristics such as sample size, species prevalence or environmental predictors have to be carefully taken into account because they can influence the models' performance (Dormann et al., 2008; Luoto et al., 2006). Furthermore, predictive models are sensitive to parameterization and selection criteria during the modelling process (Araújo and Guisan, 2006; Elith et al., 2006), and thus can result in an uncertainty of current or past/future projections of species distributions (Buisson et al., 2010; Svenning et al., 2008).

In this context, a model that can provide better insights is of prime value for assessment and management planning. Provided that the eradications of successful invaders is hardly possible (Regan et al., 2006), assessment and control of invasive/alien species are thus an urgent need to maintain biodiversity. A long-term investigation on the invasion process/success requires a rigorous analysis, which is mostly based on statistical models and their interpretation (Blossey, 1999). As most of the predictive models show varied performance when predicting species occurrence (Buisson et al., 2010; Svenning et al., 2008), improvement of the reliability of a predictive model via testing it through "hindcasting" –predicting what happened in the past – is subsequently recommended (Pearman et al., 2008; Sanders, 2012). Only when a model can provide a clear understanding of what is likely to take place in the ecosystem, managers and decision makers can construct policies to mitigate unwanted impacts. Therefore, an optimization of a particular predictive model, which can be robust and yields a relatively stable performance, will provide a better insight for further research conducted to confidently predict species distributions.

In this study, we optimized the reliability of a predictive algorithm (i.e. Classification Tree) in hindcasting and forecasting alien mollusc occurrence in river systems in Flanders, Belgium. We built 144 models based on a dataset of 20 years (1991-2010) comprising 7695 field data points, and 144 models based on a combined dataset, which is referred to the combination of the field data and cloned data (i.e. a dataset obtained by independently duplicating the filed data by *k* different individuals). In the calibration and validation process, we used random and stratified resampling approaches to split the data into different folds and set different levels of the models' complex parameters. We then evaluated how the two types of the dataset (field and combined data) and the two resampling approaches, under same and different parameterization settings, affect the model optimization. We expected that the models which were calibrated and validated using the stratified-split data are more robust and yield a greater reliability, compared to those being calibrated and validated using the random-split data, as has been shown in Hirzel & Guisan (2002).

MATERIAL AND METHODS

Data collection and processing

The Flemish Environment Agency (VVM) has been collecting biological and environmental data in Flanders since 1989. The samples were collected at more than 2500 sites spread over different water bodies. For detailed information on how the data were collected, we refer to Sor et al. (2016). Data from 1991 to 2010 was used in the present study. This dataset was then divided into 4 periods; period D1: data from 1991 to 1995, D2: 1996 to 2000, D3: 2001 to 2005 and D4: 2006 to 2010. The response variable of each site was grouped as a categorical variable: a "native" site (i.e. a site having only the native taxa present), an "alien" site (i.e. a site having only the alien taxa present) and an "overlap" site (i.e. a site having both the native and alien mollusc taxa present). Since the number of occurrence instances of most alien mollusc genera is limited (Table 1), we decided to group them into one categorical variable. In this way, we had a higher number of instances for our predictive models and thus a better and more robust model that could be developed. We removed physical and chemical water quality variables containing missing values in more than 5% of all samples. Finally, each period consisted of one categorical response variable (native/alien/overlap) and 11 environmental predictors, which are summarized in Table 2.

Modelling algorithm

We used a Classification Tree (CT) model to predict the occurrence of the response variable. The CT is known as an easy-to-read and easy-to-interpret method (Guisan and Thuiller, 2005). Due to this simplicity and interpretability, and its ability to tolerate missing values during the calibration and validation process (De'ath and Fabricius, 2000; Therneau and Atkinson, 1997), CT is mainly implemented to model complex biological interactions and is suggested for management applications (Guisan and Thuiller, 2005; Waite et al., 2010). For a detailed description of the theoretical growing and pruning process of the CT models, we refer to Quinlan (1986) and Therneau & Atkinson (1997).

Hindcasting, forecasting and validation

Field data

For the hindcasting**,** we only used data period D4 to calibrate the models. Then the calibrated models were used to hindcast (validate) the response variable based on the environmental predictors of the data from period D4, D3, D2 and D1. During the pruning phase, the complex parameter of the calibrated models was set at cp=0.05. Where the tree had only a root, we decreased the cp to a level (e.g. cp between 0.04 - 0.01) that at least two terminal nodes were produced. The first few splits usually provide a very informative division of the data (Therneau and Atkinson, 1997). These criteria were set in order to make the trees easily interpretable and comparable in terms of the number of variables and complexity.

For the forecasting, we only used data period D1 to calibrate the models. Then we used the calibrated models to forecast (validate) the response variable based on the environmental predictors of the data from period D1, D2, D3 and D4. The setting of the complex parameter of the calibrated models was similar to the settings as described in the "hindcasting" section.

The calibration and validation process was based on a three-fold cross-validation (3CV). The 3 fold CV was obtained by splitting the whole dataset into three folds. Two folds of the 3 fold CV were in turn used for the calibration, whereas one fold was used for the validation. In other words, we used two-thirds to calibrate and one-third to validate the models, and repeated the procedure three times so that at the end each fold was used exactly once for the validation (hindcasting and forecasting). For the data splitting, we used a random and a stratified resampling approach.

The random resampling was done by shuffling and randomly splitting the whole dataset into three folds. From this approach, the sample size of each class was not equally stratified across the three folds. To allow a reliable error estimation of the hindcasting and forecasting models for each period, we made three replicates of the 3 fold CV procedure by reshuffling and randomly splitting the whole dataset again into 3 new folds.

For the stratified resampling, we first allocated the three classes of the response variable and the environmental predictors of the whole dataset into three separated data subsets (DS, i.e. DS1, DS2, DS3). Each DS corresponds to the data of each class of the response variable. Then, we randomly divided each DS into 3 smaller data subsets (i.e. DS1a, DS1b, DS1c; DS2a, DS2b, DS2c; DS3a, DS3b, DS3c). Thus, each of the smaller data subsets of the same class had exactly the same number of samples (n, e.g. nDS1a = nDS1b = nDS1c). For any DS that could not be equally divided by 3, each of the smaller data subsets may have 1 sample more or less than the others (e.g. nDS1a ± 1). Finally, we recombined each smaller data subset to obtain new three equal-stratified data subsets, i.e. Fold1 (nDS1a + nDS2a + nDS3a), Fold2 (nDS1b + nDS2b + nDS3b) and Fold3 (nDS1c + nDS2c + nDS3c), which were later used to calibrate and validate the models. The schematic diagram showing this resampling procedure is depicted in Fig. 1. To be consistent, we also made three replicates of the 3 fold CV for this resampling approach, by reshuffling the original data and then following the above-described procedure. The calibration and validation of the models was performed separately for each resampling approach.

Combined data

In most cases, the "native" class had the largest sample size, followed by the "overlap" and the "alien" class. Due to the unequal sample size, we developed a cloned dataset by independently duplicating the original filed data of the less frequently occurring classes by *k* different individuals. Then we combined the cloned data with the original field data to obtain the same sample size for each class (Fig. 2). One may argue to use an equal-stratified dataset of field observations by randomly selecting a number of observation of the more occurring classes equally to the less occurring class. If following this, however, we will lose incredible valuable-information of the field observations and that leaves a small sample size for the model development, which consequently leads to a less reliable and less robust models. On the contrary, using the cloned data allows us to incorporate all field observations and provides us an equal data distribution and sufficient samples for each period. Incorporating the cloned data into the models will increase the maximum likelihood of the prior class distribution, making the models more robust (Lele et al., 2007). To run the hindcasting, forecasting and validations for each period using the combined data, we followed the criteria and procedures used in the "field data" section.

Model performance and analysis

For each scenario (hindcasting and forecasting), nine models (3 models of 3 fold CV × 3 replicates) were built for each resampling approach in each period (see Supporting Information Table S1). In this way, a performance value of the nine different models was calculated. From the field data, we built 36 models ([3 models of 3 fold CV hindcasting × 3 replicates × 2 resampling types] + [3 models of 3 fold CV forecasting × 3 replicates × 2 sampling types]) for each period. In the same way, 36 models were built based on the combined data for each period. Therefore, we totally built 144 models over the four periods for each data type (field and combined data).

Cohen's Kappa Statistic (K) and Correctly Classified Instances (CCI) were used to evaluate the model performance. The higher the value of K (ranging from 0 to 1) and of CCI (ranging from 0 to 100), the better the model predicts the response variable. To obtain the best model configuration in both scenarios, we analysed the model performance based on 1) the data types, 2) the resampling approaches and 3) the data types and resampling approaches. All the analyses were carried out using the language program R (R Core Team, 2013).

RESULTS

Field data

For the hindcasting scenario, the yielded predictive performance (i.e. Kappa and CCI) of the CT models were Kappa = 0.16, 0.16, 0.11 and 0.07, and CCI = 55%, 58%, 60% and 50% for the data period D4, D3, D2 and D1 (Fig. 3). The yielded Kappa and CCI for the forecasting scenario were 0.35, 0.07, 0.07 and 0.03, and 93%, 91%, 79% and 47%, respectively. The performance of the models, which were calibrated and validated using the two resampling approaches, showed a relatively similar trend and stability (see Supporting information Fig. S1). However, the CT models were not able to hindcast and forecast the class "alien", although they could hindcast and forecast the other two classes (see Table S2).

Combined data

The Kappa (0.24, 0.16, 0.13, 0.02) and overall CCI (48%, 44%, 42%, 24%) of the models decreased from period D4, D3, D2 to D1 for the hindcasting scenario. For the forecasting scenario, the model performance decreased from period D1 to D4; the Kappa was 0.58, 0.17, -0.01 and -0.05, and overall CCI was 72%, 44%, 33% and 30% for the corresponding periods (Fig. 4a-b). The calibrated and validated models using the two resampling approaches also yielded a similar trend in predictive performance and in stability for each scenario (see Fig. S1). Interestingly, the models were able to hindcast and forecast the class "alien" and the other two classes. The CCI of the class "alien" followed the decreasing trend of the overall performance for each scenario; it was 76%, 62%, 44% and 9%, and 100%, 44%, 20% and 14% for the corresponding periods used in the hind- and forecasting scenario, respectively (Fig. 4c).

DISCUSSION

In this study, we optimized the hind- and forecasting of alien mollusc occurrence to obtain the best CT model configuration for both in terms of performance and interpretation. Variation in class distribution and resampling design were investigated and strongly indicated that a similar sample size of each class should be considered as has been generally suggested (e.g. Manel *et al.* 2001; Allouche *et al.* 2006). Based on field data (field observations), for which the prevalence of each predicted class was not equally distributed, the CT models were unable to predict the putative occurrence of the alien molluscs. These results indicate that the field observations could not be predicted. However, by using the combined data, their occurrence could be correctly predicted at periods other than when the CT models were calibrated, which corresponds to the field past and current situation.

Improving the model reliability by testing it to hindcast the occurrence of a given species in the past could be decisive for predicting the future occurrence of that species. Although predictive models have been substantially implemented in ecology in the last two decades (Guisan et al., 2013; Guisan and Thuiller, 2005), only a few have tested each model's performance via hindcasting (e.g. Svenning *et al.* 2008; Espíndola *et al.* 2012; Maire *et al.* 2015; Pelletier *et al.* 2015). Most of the existing studies have used currently known species occurrences to hindcast or forecast the past or future putative species occurrence (Boets et al., 2013; Buisson et al., 2010; Maire et al., 2015; Svenning et al., 2008), which sometimes results in a high uncertainty of the models used. In the present study, we used both the hindcasting- and forecasting-based models to predict the occurrence alien molluscs. As a result CT models could make a correct prediction of the past and recent occurrence.

One of the challenges in predicting a species' occurrence is obtaining a balanced class distribution for the response variable, because a small sample size of each predicted class can be the source of instability and errors in species distribution models (Allouche et al., 2006; McPherson et al., 2004). However, our results demonstrated that having the exact same class distribution (i.e. stratified split) is not always necessary during the calibration and validation process because the models can make the correct prediction and yield a similar performance when the prevalence reaches a certain threshold (see Fig. S1). However, the model is incapable of predicting a class with a very low prevalence, as was the case of the class "alien" in this study (see Table S2). As collecting new field data from past/recent periods is unfeasible/costly, optimization appears necessary (Hirzel and Guisan, 2002). In respond to this, we optimized the prediction by considering cloned data for the models, resulting in a better prediction. Cloned data have been applied and recently suggested for hierarchical models in ecology (Lele et al., 2010, 2007; Ponciano et al., 2009). The increasing number of clones used in the models can increase the maximum likelihood of the prior class distribution, but it does not affect the statistical accuracy, which mainly depends on the information of the field data and the model calibration and validation process (Lele et al., 2007).

Parameterization and selection criteria during the modelling process are known to be sensitive to model performance (Araújo and Guisan, 2006; Elith and Graham, 2009). In this study, the complex parameter (cp) ranging from 0.05 to 0.01 was used to optimize the tree selection of the calibrated models. In most cases, this setting resulted in a tree having 2 to 4 nodes, with an increased coefficient of determination (R^2) of the model cross-validation (i.e. reducing cross-validation error) (see Fig. S2). The trees having a few nodes and yielding an increased R^2 are known to provide a good reliability and representation of data division (Therneau et al., 2015; Therneau and Atkinson, 1997). This suggests that the complexity parameter and selection criteria set in our study may not lead to over-fitting of the hind- and forecasted models.

In the best model configuration (using the combined data), environmental variables, which are considered as important factors promoting the occurrence and abundance of alien molluscs, were incorporated in the calibrated and validated models. They include water quality variables such as nitrate, ammonium, conductivity, pH, and chemical oxygen demand (Boets et al., 2013; Grabowski et al., 2009; Strayer, 2010; Vermonden et al., 2010). The water quality conditions in Flanders, which have been substantially improved over the past decades (MIRA, 2012), could be the factors leading to a gradual decrease in global performance of both the hind- and forecasting models in our study. This result might also reveal the natural characteristics of predictive models that the uncertainty increases when longer predictions of the future are made (Buisson et al., 2010; Pearman et al., 2008). Furthermore, alien species are mostly considered generalists, being able to withstand a wide range of environmental conditions (Werner & Rothhaupt 2007; Pigneur *et al.* 2014). This characteristic, along with their currently increased migration patterns and the changes in water quality conditions, could influence the correct prediction of the class “alien” when predicting the past or to the present (Fig. 1). Nevertheless, although the CCI decreased from recent to past periods (hindcasting scenario) and vice-versa (forecasting scenario), the CT models could still correctly predict the occurrence of alien molluscs with a low error rate (Fig. 4), demonstrating the robustness of the CT models.

While we have incorporated all important environmental factors to optimize the CT models, available data of dispersal vectors might provide a better model output. Although these vectors are not always a limiting factors, previous studies have shown its contribution in predicting the distribution of alien species (Boets et al., 2014, 2013). Regardless the dispersal vectors, however, if the alien species maintain their climatic niche, the CT models will be able to predict their future occurrence on the basis of the current environment conditions in which they have already established (Parravicini et al., 2015). Otherwise, climatic variables should be considered in the models, because such variables, e.g. temperature, have been illustrated to associate with the establishment success of alien species (Leuven et al., 2009; Werner and Rothhaupt, 2008). Consideration of incorporating these variables and applying the hind-- and forecasting-based approach outside Flanders will also help to test the robustness of CT models and to improve our understanding of the behavior and ecological characteristics of alien molluscs, which are imperative for assessment and management purposes.

CONCLUSION

Since the rate at which studies employing predictive models increases and that their performance uncertainty is always noticed, evaluating the reliability of a particular model via hindcasting is a proposition for further studies aiming to predict future species occurrence. Here we report the robustness of CT models in correctly predicting alien mollusc occurrence by incorporating cloned data into the models. With the same parameterization, the performance of the CT models is unlikely to be affected by resampling approaches (i.e. random vs stratified) used to split the data for calibrating and validating the models.

Our optimization success provides insight for the application of predictive models. Although model reliability decreased when predicting further into the past or into the future, the models still provide a correct prediction with a low error rate. The CT model is therefore suitable for assessment and management applications. Incorporating dispersal vectors and climatic variables into the models and testing model transferability into a new area may also help to increase the robustness of the CT models.

ACKNOWLEDGEMENTS

We would like to thank the Flemish Environment Agency (VMM) for the opportunity to study their samples. RS is supported by the Erasmus Mundus LOTUS (European Commission) and the Forum Belmont TLSCC project. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

AUTHORSHIP

RS, PB and PG conceived the study, RS and SL performed the modelling work, RS wrote the first draft of the manuscript, and all authors contributed significantly to revisions.

REFERENCES

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Bernauer, D., Jansen, W., 2006. Recent invasions of alien macroinvertebrates and loss of native species in the upper Rhine River, Germany. Aquat. Invasions 1, 55–71.

Blossey, B., 1999. Before, during and after: the need for long-term monitoring in invasive plant species management. Biol. Invasions 1, 301–311.

Boets, P., Lock, K., Goethals, P.L.M., 2013. Modelling habitat preference, abundance and species richness of alien macrocrustaceans in surface waters in Flanders (Belgium) using decision trees. Ecol. Inform. 17, 73–81.

Boets, P., Pauwels, I., Lock, K., Goethals, P., 2014. Using an integrated modelling approach for risk assessment of the "killer shrimp" Dikerogammarus villosus. River Res. Appl. 30, 403–412. doi:DOI: 10.1002/rra.2658

Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Chang. Biol. 16, 1145–1157.

Chen, L., Peng, S., Yang, B., 2015. Predicting alien herb invasion with machine learning models: biogeographical and life-history traits both matter. Biol. Invasions 17, 2187–2198.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

Didham, R.K., Tylianakis, J.M., Gemmell, N.J., Rand, T., Ewers, R., 2007. Interactive effects of habitat modification and species invasion on native species decline. Trends Ecol. Evol. 22, 489–96.

Dormann, C.F., Purschke, O., Garcia Marquez, J.R., Lautenbach, S., Schroder, B., 2008. Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. Ecology 89, 3371–3386.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., Naiman, R.J., Prieur-Richard, A.-H., Soto, D., Stiassny, M.L.J., Sullivan, C.A., 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol. Rev. Camb. Philos. Soc. 81, 163–182.

Elith, J., Graham, C.H., 2009. Do they? How do they? Why do they differ? on finding reasons for differing performances of species distribution models. Ecography (Cop.). 32, 66–77.

Elith, J., Graham, C.H., Anderson, R.P., Ferrier, S., Dudík, M., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC Overton, J., Peterson, T.A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop.). 29, 129–151.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.

Espíndola, A., Pellissier, L., Maiorano, L., Hordijk, W., Guisan, A., Alvarez, N., 2012. Predicting present and future intra-specific genetic structure through niche hindcasting across 24 millennia. Ecol. Lett. 15, 649–657.

Evangelista, P.H., Kumar, S., Stohlgren, T.J., Jarnevich, C.S., Crall, A.W., Norman, J.B., Barnett, D.T., 2008. Modelling invasion for a habitat generalist and a specialist plant species. Divers. Distrib. 14, 808–817.

Grabowski, M., Bacela, K., Konopacka, A., Jazdzewski, K., 2009. Salinity-related distribution of alien amphipods in rivers provides refugia for native species. Biol. Invasions 11, 2107–2117.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., Mcdonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., Peterson, A.T., 2007. What matters for predicting the occurrences of trees: techniques, data, or species characteristics? Ecol. Monogr. 77, 615–630.

Gurevitch, J., Padilla, D.K., 2004. Are invasive species a major cause of extinctions? Trends Ecol. Evol. 19, 470–474.

Hill, M.P., Chown, S.L., Hoffmann, A.A., 2013. A predicted niche shift corresponds with increased thermal resistance in an invasive mite, Halotydeus destructor. Glob. Ecol. Biogeogr. 22, 942–951.

Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecol. Modell. 157, 331–341.

Lele, S.R., Dennis, B., Lutscher, F., 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol. Lett. 10, 551–563.

Lele, S.R., Nadeem, K., Schmuland, B., 2010. Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. J. Am. Stat. Assoc. 105, 1617–1625.

Leuven, R.S.E.W., van der Velde, G., Baijens, I., Snijders, J., van der Zwart, C., Lenders, H., bij de Vaate, A., 2009. The river Rhine: a global highway for dispersal of aquatic invasive species. Biol. Invasions 11, 1989–2008.

Luoto, M., Heikkinen, R.K., Pöyry, J., Saarinen, K., 2006. Determinants of the biogeographical distribution of butterflies in boreal regions. J. Biogeogr. 33, 1764–1778.

Maire, A., Buisson, L., Canal, J., Rigault, B., Boucault, J., Laffaille, P., 2015. Hindcasting modelling for restoration and conservation planning: Application to stream fish assemblages. Aquat. Conserv. Mar. Freshw. Ecosyst. 25, 839–854.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology; the need to count for prevalence. J. Appl. Ecol. 38, 921–931.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.

MIRA, 2012. Flanders Environment Report - Indicator Report.

Parravicini, V., Azzurro, E., Kulbicki, M., Belmaker, J., 2015. Niche shift can impair the ability to predict invasion risk in the marine realm: An illustration using Mediterranean fish invaders. Ecol. Lett. 18, 246–253.

Pearman, P.B., Randin, C.F., Broennimann, O., Vittoz, P., Knaap, W.O. Van Der, Engler, R., Lay, G. Le, Zimmermann, N.E., Guisan, A., 2008. Prediction of plant species distributions across six millennia. Ecol. Lett. 11, 357–369.

Pelletier, T.A., Crisafulli, C., Wagner, S., Zellmer, A.J., Carstens, B.C., 2015. Historical species distribution models predict species limits in western plethodon salamanders. Syst. Biol. 64, 909–925.

Pigneur, L.-M., Falisse, E., Roland, K., Everbecq, E., Deliège, J.-F., Smitz, J., Van Doninck, K., Descy, J.-P., 2014. Impact of invasive Asian clams, Corbicula spp., on a large river ecosystem. Freshw. Biol. 59, 573–583.

Ponciano, J.M., Taper, M.L., Dennis, B., Lele, S., 2009. Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. Ecology 90, 356–362.

Quinlan, J.R., 1986. Induction of Decision Trees. Mach. Learn. 1, 81–106.

R Core Team, 2013. R: a language and environment for statistical computing.

Regan, T.J., McCarthy, M.A., Baxter, P.W.J., Dane Panetta, F., Possingham, H., 2006. Optimal eradication: when to stop looking for an invasive plant. Ecol. Lett. 9, 759–766.

Richter, B.D., Braun, D.P., Mendelson, M. a, Master, L.L., 1997. Threats to imperilled freshwater fauna. Conserv. Biol. 11, 1081–1093.

Roura-Pascual, N., Brotons, L., Peterson, A.T., Thuiller, W., 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biol. Invasions 11, 1017–1031.

Sala, O.E., Chapin III, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L., Jackson, R., Kinzig, A., Leemans, R., Lodge, D., Mooney, H., Oesterheld, M., Poff, N., Skykes, M., Walker, B., Walker, M., Wall, D., 2000. Global biodiversity scenarios for the year 2100. Science (80-.). 287, 1770–1774.

Sanders, R., 2012. Hindcasting helps scientists improve forecasts for life on Earth. UC Berkeley News.

Segurado, P., Araujo, M., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.

Strayer, D.L., 2010. Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. Freshw. Biol. 55, 152–174.

Svenning, J.C., Normand, S., Kageyama, M., 2008. Glacial refugia of temperate trees in Europe: insights from species distribution modelling. J. Ecol. 96, 1117–1127.

Therneau, T., Atkinson, B., Ripley, B., Ripley, M.B., 2015. Package “rpart.”

Therneau, T.M., Atkinson, E.J., 1997. An introduction to recursive partitioning using the rpart routines. Technical report no. 61. Rochester, Minnesota.

Thuiller, W., Richardson, D.M., Py Ek, P., Midgley, G.F., Hughes, G.O., Rouget, M., 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Glob. Chang. Biol. 11, 2234–2250.

Vermonden, K., Leuven, R.S.E.W., Van Der Velde, G., 2010. Environmental factors determining invasibility of urban waters for exotic macroinvertebrates. Divers. Distrib. 16, 1009–1021.

Vicente, J., Randin, C.F., Gonçalves, J., Metzger, M.J., Lomba, Â., Honrado, J., Guisan, A., 2011. Where will conflicts between alien and rare species occur after climate and land-use change? a test with a novel combined modelling approach. Biol. Invasions 13, 1209–1227.

Waite, I.R., Brown, L.R., Kennen, J.G., May, J.T., Cuffney, T.F., Orlando, J.L., Jones, K. a., 2010. Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. Ecol. Indic. 10, 1125–1136.

Werner, S., Rothhaupt, K.-O., 2007. Effects of the invasive bivalve Corbicula fluminea on settling juveniles and other benthic taxa. J. North Am. Benthol. Soc. 26, 673–680.

Werner, S., Rothhaupt, K.O., 2008. Mass mortality of the invasive bivalve Corbicula fluminea induced by a severe low-water event and associated low water temperatures. Hydrobiologia 613, 143–150.

List of tables and figures with captions

Table 1. The occurrence instances of each alien mollusc genus and of all genera that are merged together, compared to all collected samples (7695 samples) within the studied period (1991-2010).

Genus	*Corbicula*	*Dreissena*	*Ferrisia*	*Lithoglyphus*	*Menetus*	*Physella*	*Viviparus*	Grouping all genera
Instances	130	745	381	30	4	1138	94	2028
Instances (%)	1.4	8.8	4.5	0.4	<0.1	13.9	1.1	26.4

Table 2. Mean value (and standard deviation) for environmental predictors, and occurrence instances of each class of the response variable. NH_4^+: Ammonium, COD: Chemical Oxygen Demand, Pt: Total Phosphorus, EC: Electrical Conductivity, NO_3^-: Nitrate, NO_2^-: Nitrite, oPO_4: Orthophosphate, DO: Dissolved Oxygen, SI: Sinuosity.

		Periods			
Variable	Unit	D1: 1991-1995	D2: 1996-2000	D3: 2001-2005	D4: 2006-2010
NH_4^+	mg/L	2.6 (4.8)	1.9 (3.4)	1.6 (2.8)	2.3 (5.1)
COD	mg/L	55 (43)	39 (42)	34 (34)	36 (41)
Pt	mg/L	1.0 (1.4)	0.9 (1.4)	0.9 (2.3)	0.8 (1.0)
EC	µS/cm	1320 (2618)	987 (1149)	998 (1509)	921 (949)
NO_3^-	mg/L	3.3 (4.1)	4.0 (4.7)	3.6 (3.7)	3.0 (3)
NO_2^-	mg/L	0.2 (0.3)	0.2 (0.3)	0.2 (0.2)	0.2 (0.2)
oPO_4	mg/L	0.6 (0.9)	0.5 (1.1)	0.4 (0.7)	0.5 (0.8)
pH		7.5 (0.6)	7.6 (0.5)	7.7 (0.4)	7.6 (0.4)
DO	mg/L	7.5 (3.5)	6.8 (3)	6.9 (3.1)	6.6 (2.9)
SI		1.1 (0.1)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)
Slope	m/1000m	1.0 (1.4)	1.6 (2.8)	2.0 (3.6)	1.7 (2.8)
Response	class				
Native		*863*	*1860*	*2259*	*842*
Alien		*6*	*15*	*63*	*133*
Overlap		*66*	*146*	*567*	*875*
Total instances		935	2021	2889	1850

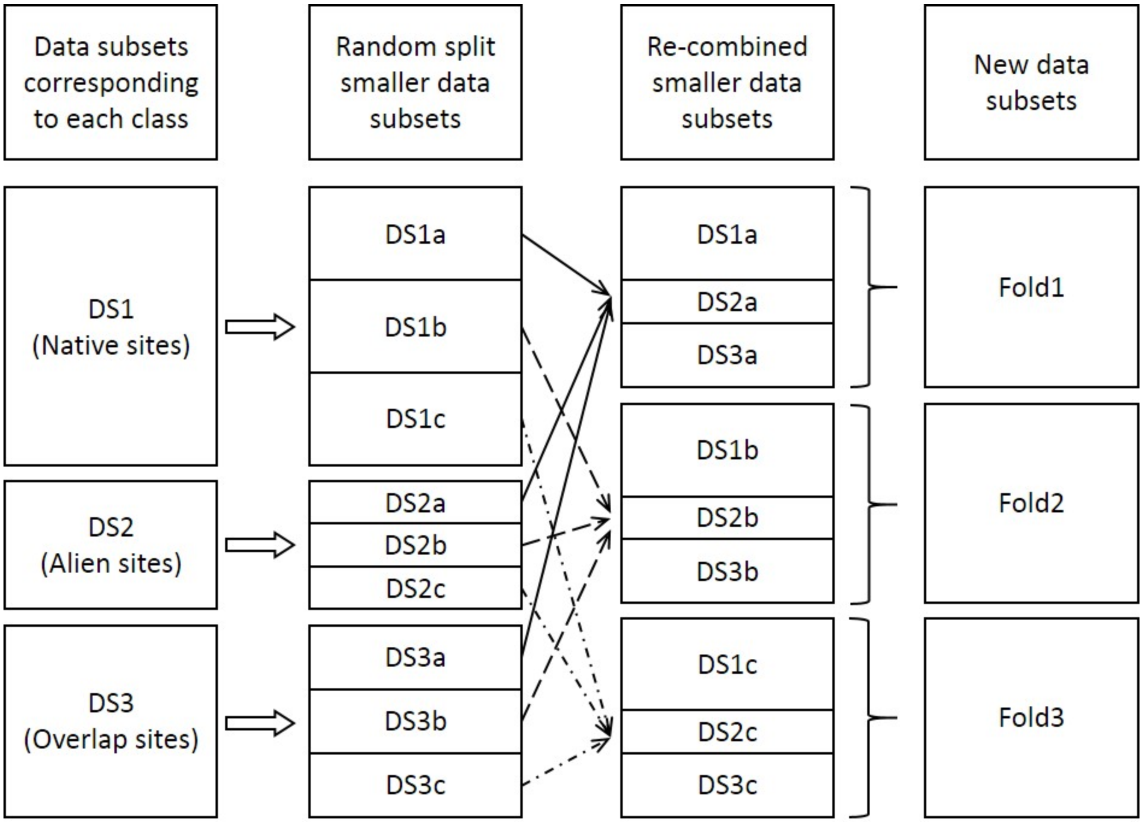


Fig. 1 Schematic diagram illustrating the splitting procedure of stratified resampling approach.

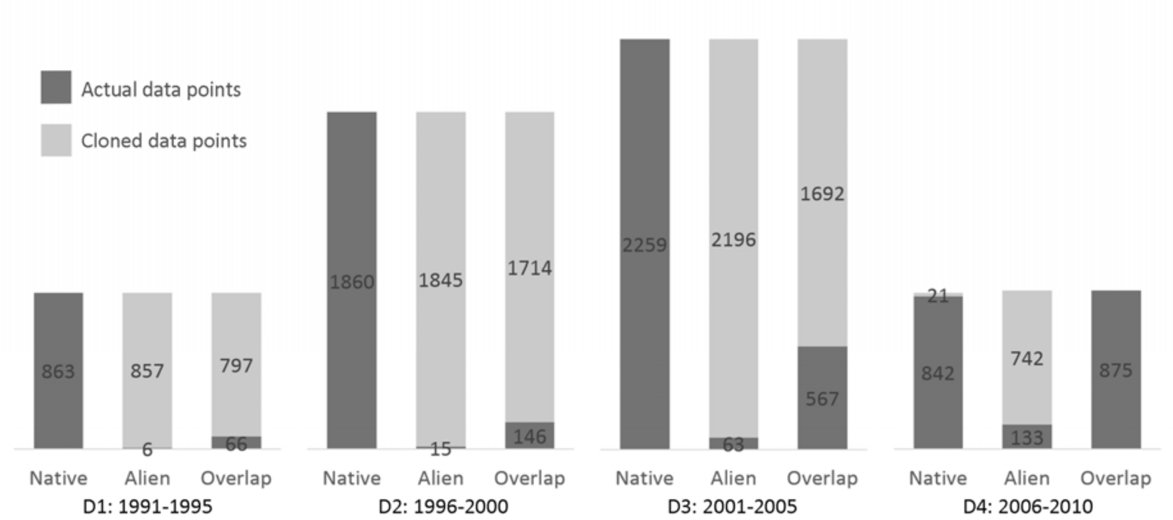


Fig. 2 Bar charts showing the number of the field and cloned data, which were combined together and used in the models.

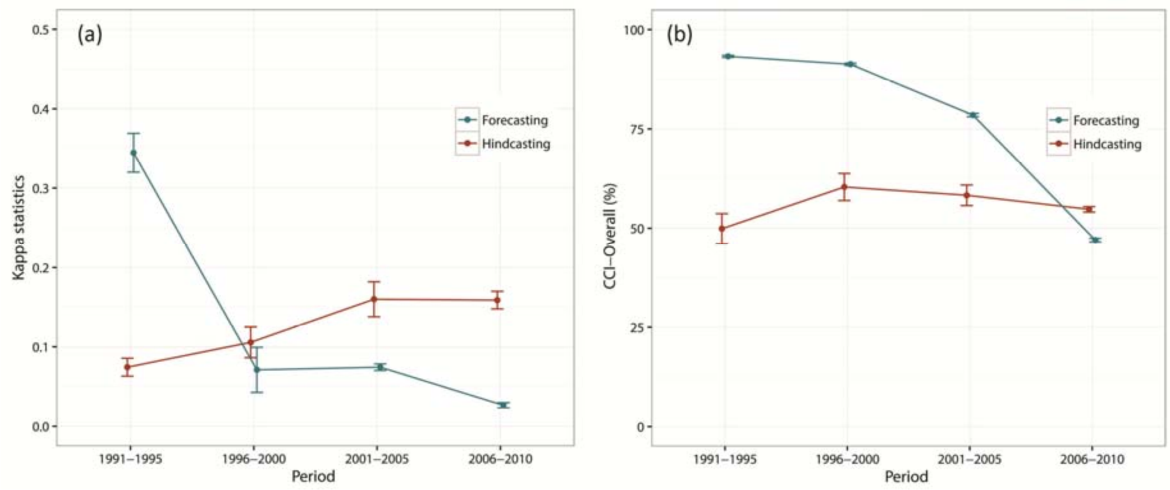


Fig. 3 Mean and standard error bar indicating the overall performance of CT models based on Kappa (a) and CCI (b), when using the field data and random resampling approach.

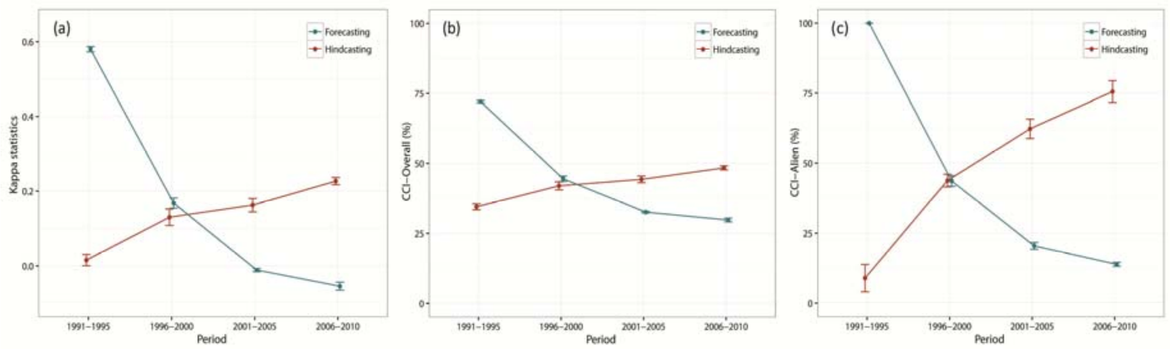


Fig. 4 Mean and standard error bar indicating the performance of CT models based on Kappa and CCI, when using combined data and random resampling approach. (a) and (b) are the overall performance, and (c) illustrates the correct prediction of the predicted class "alien".

Supporting information

Fig. S1 Mean and standard error bar indicating the overall performance of CT models based on Kappa and CCI, when using the field data (Fig. 1, 2) and combined data (Fig. 3, 4). Each column corresponds to each resampling approach (random and stratified).

Fig. S2 The 288 trees built for the two scenarios using the two types of dataset (field and combined data) and the two resampling approaches (random and stratified). The following slides, from slide number 1 to 10 and from 11 to 20, are the trees using the field and combined data, respectively. For the hindcasting scenario, the models were calibrated using the data period D4 and were validated using the environmental predictor of the data period D4, D3, D2 and D1. For the forecasting scenario, the models were calibrated using the data period D1 and were validated using the environmental predictor of the data period D1, D2, D3 and D4. The calibrated models' complex parameter and their cross-validation error (xerror) are given under each tree. The increased xerror (↑) is highlighted in yellow. The decreased and stable xerrors are respectively labeled with (↓)and (~). The detailed information of each slide title is provided in the following:

- Field_Hind_V-D4(3,2,1)_Random/Stratified: models were calibrated using the field data, for hindcasting scenario (validation based on D4, D3, D2, D1), and using random/ stratified resampling approach.
- Field_Fore_V-D1(2,3,4)_Random/Stratified: models were calibrated using the filed data, for forecasting scenario (validation based on D1, D2, D3, D4), and using random/stratified resampling approach.
- Comined_Hind_V-D4(3,2,1)_Random/Stratified: models were calibrated using the combined data, for hindcasting scenario (validation based on D4, D3, D2, D1), and using random/ stratified resampling approach.

Comined_Fore_V-D1(2,3,4)_Random/Stratified: models were calibrated using the combined data, for forecasting scenario (validation based on D1, D2, D3, D4), and using random/stratified resampling approach.

Table S1. Summary of the 288 models which were built based on the two types of dataset (field and combined data) and the two resampling approaches.

Data	Scenario	Re-sampling	Period			
			D1	D2	D3	D4
Field data	Hindcasting	R	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
		S	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
	Forecasting	R	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
		S	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
Combined data	Hindcasting	R	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
		S	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
	Forecasting	R	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep
		S	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep	3CV × 3Rep

R: random resampling, S: stratified resampling, CV: cross-validation, Rep: replicate, which refers to a different seed set chosen for the determination of new folds.

Table S2. Confusion matrices showing the instances of the observed and predicted classes obtained from the CT models of the hindcasting and forecasting scenarios for each period. The instances of each class are averaged from the 9 models which were calibrated and validated using random resampling approach. The three classes are overlap (O), Alien (A) and Native (N).

Period		**D1: 1991-1995**				**D2: 1996-2000**				**D3: 2001-2005**				**D4: 2006-2010**			
Field data																	
Hindcasting				Observed				Observed				Observed				Observed	
			O	A	N		O	A	N		O	A	N		O	A	N
	Predicted	O	18	2	150	O	34	4	247	O	126	13	318	O	152	22	94
		A	0	0	0	A	0	0	0	A	0	0	0	A	1	0	1
		N	4	0	138	N	15	1	373	N	63	8	435	N	139	22	186
Forecasting				Observed				Observed				Observed				Observed	
			O	A	N		O	A	N		O	A	N		O	A	N
	Predicted	O	6	0	3	O	3	0	8	O	13	0	10	O	13	1	4
		A	0	0	0	A	0	0	0	A	0	0	0	A	0	0	0
		N	16	2	285	N	46	5	612	N	176	21	743	N	279	44	277
Combined data																	
Hindcasting				Observed				Observed				Observed				Observed	
			O	A	N		O	A	N		O	A	N		O	A	N
	Predicted	O	183	221	62	O	289	263	135	O	310	203	202	O	94	51	75
		A	61	26	135	A	166	271	265	A	314	472	338	A	124	220	108
		N	43	43	88	N	165	86	220	N	121	81	218	N	73	21	109
Forecasting				Observed				Observed				Observed				Observed	
			O	A	N		O	A	N		O	A	N		O	A	N
	Predicted	O	154	0	64	O	153	83	90	O	114	52	133	O	33	15	37
		A	43	288	44	A	275	270	126	A	259	154	153	A	72	41	67
		N	91	0	180	N	192	267	404	N	380	547	467	N	186	236	187

Supporting Information

Figure S1 Mean and standard error bar indicating the overall performance of CT models based on Kappa and CCI, when using the field data (Fig. 1, 2) and combined data (Fig. 3, 4). Each column corresponds to each resampling approach (random and stratified).

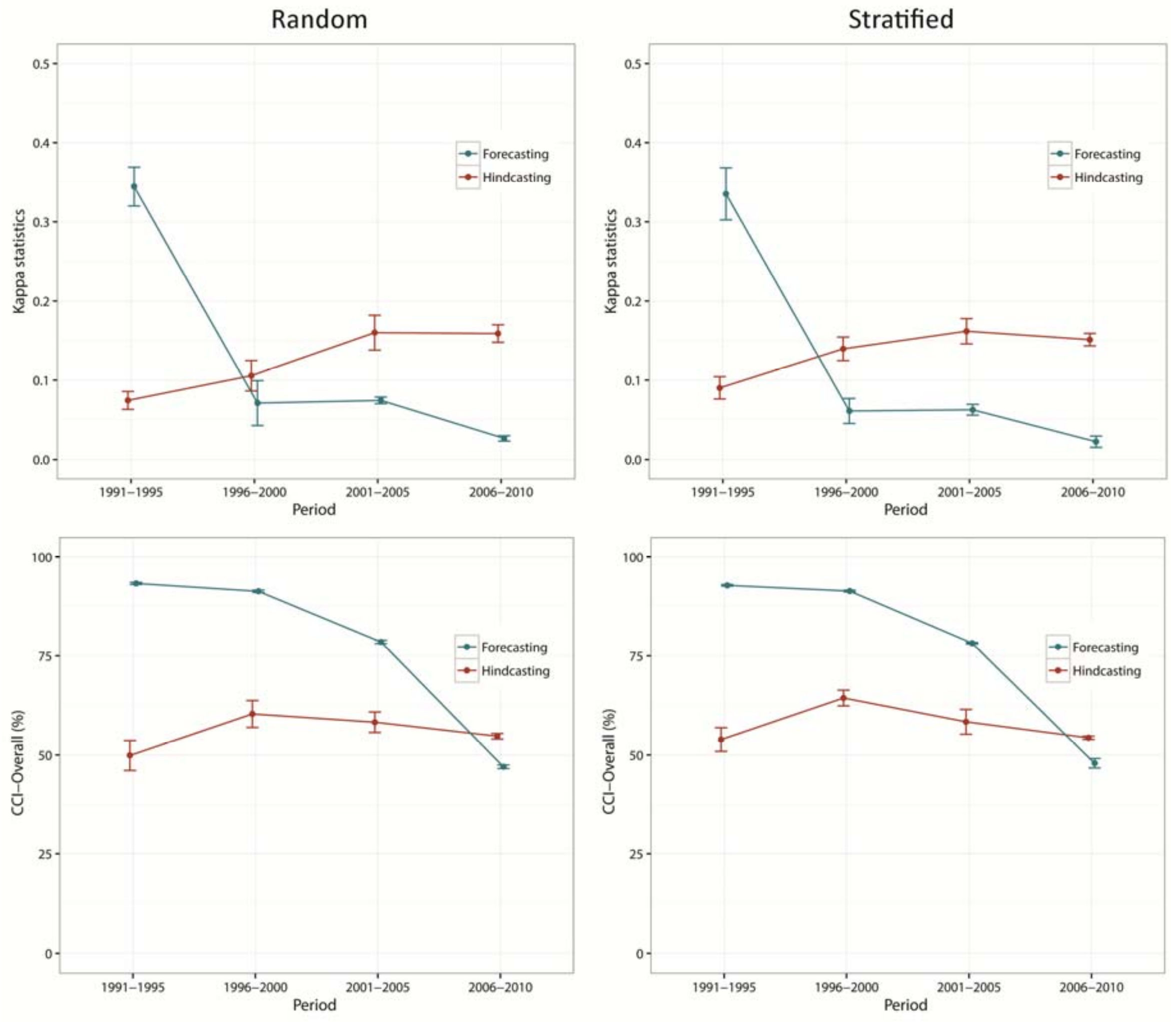


Fig. 1 The overall performance of models using the field data.

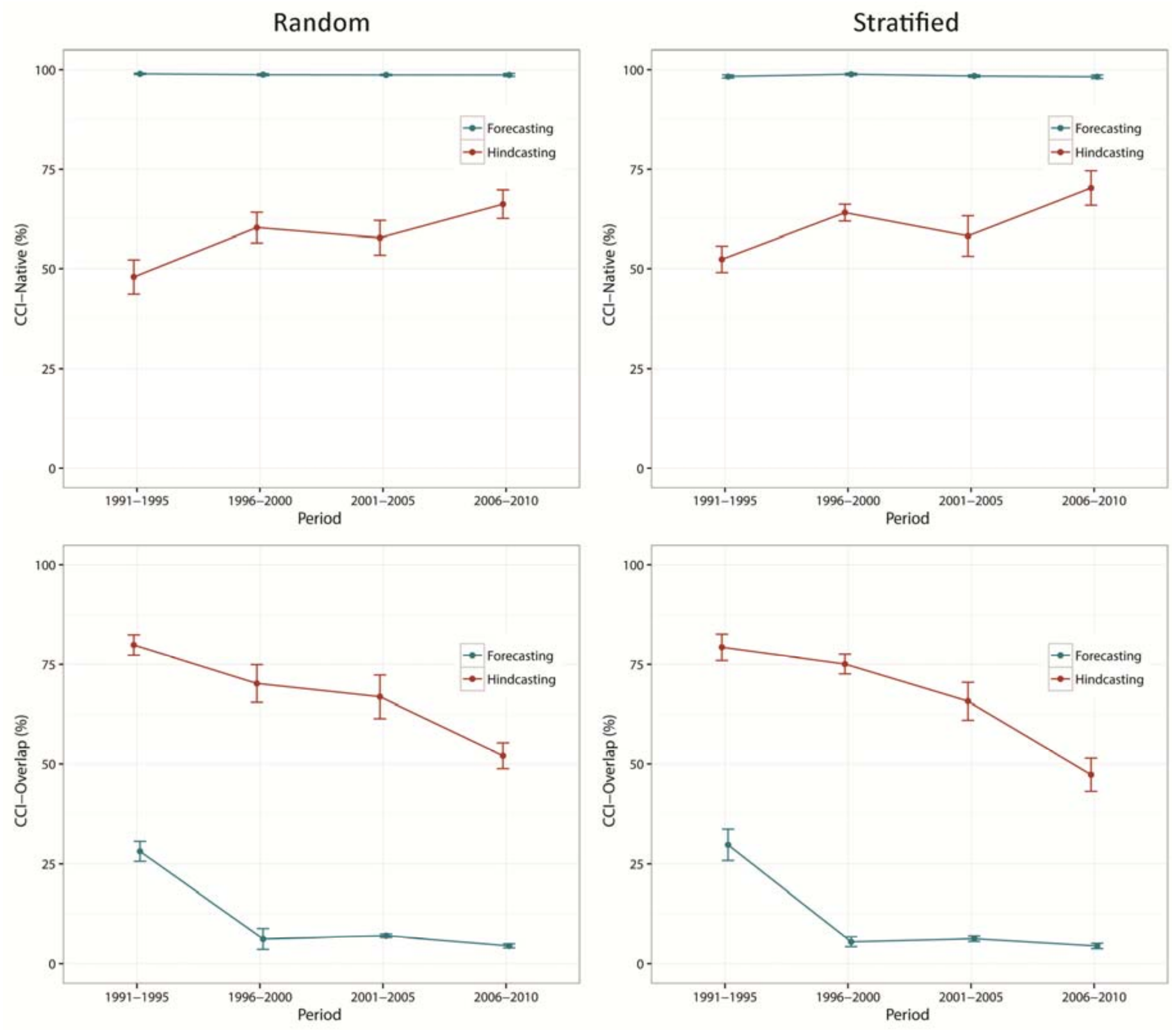


Fig. 2 The CCI of the predicted classes of models using the field data.

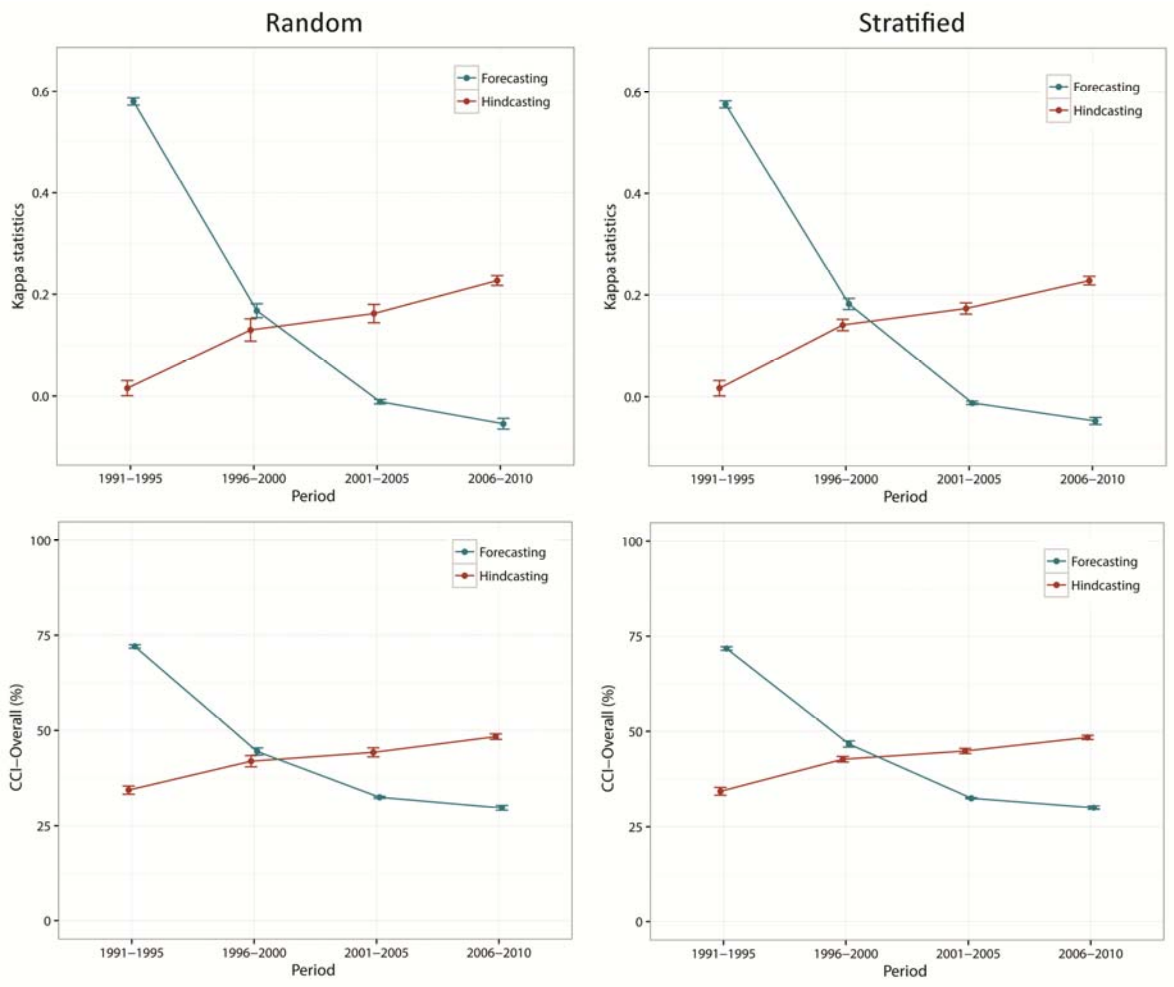


Fig. 3 The overall performance of models using the combined data.

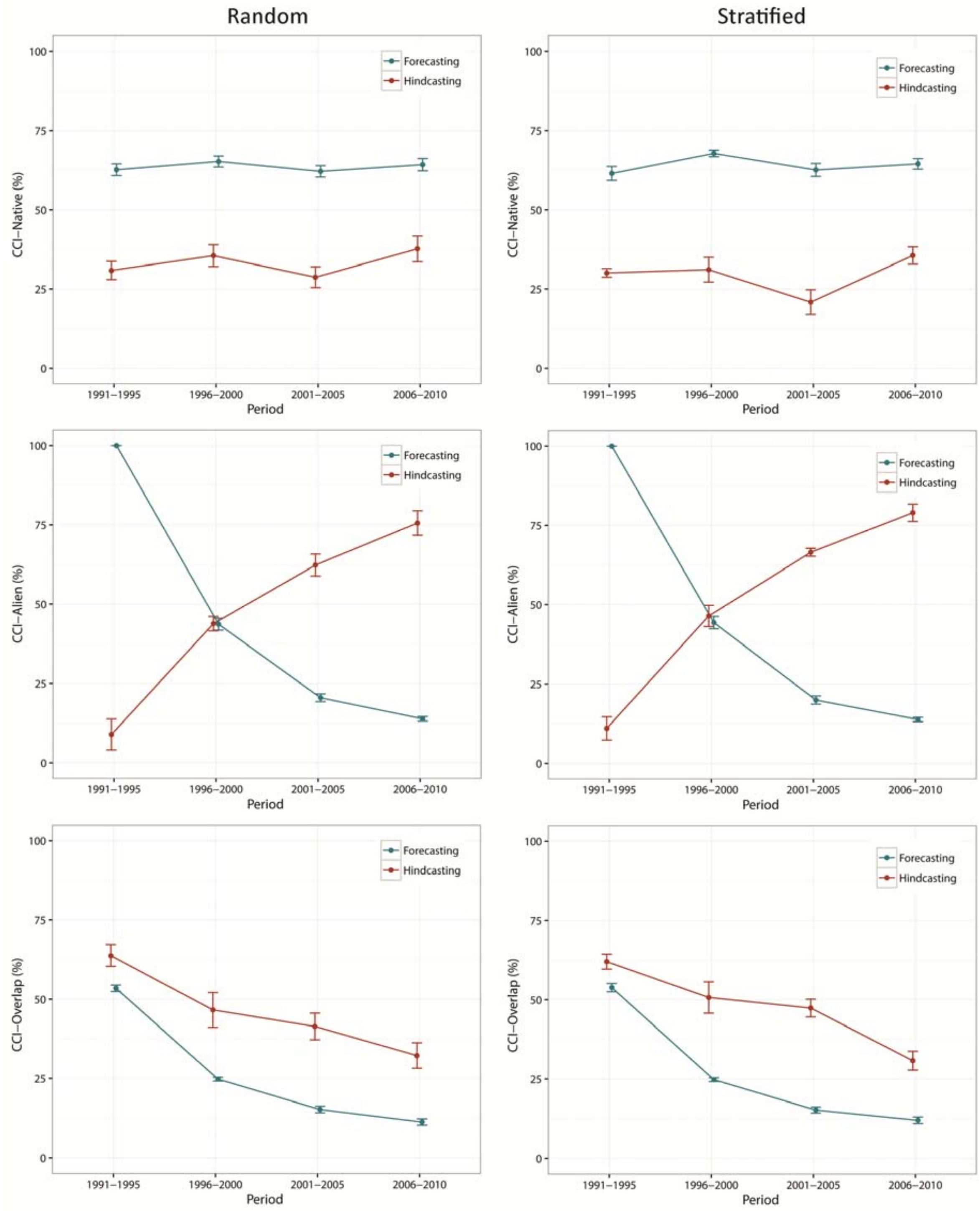


Fig. 4 The CCI of the predicted classes of models using the combined data.

Supporting Information

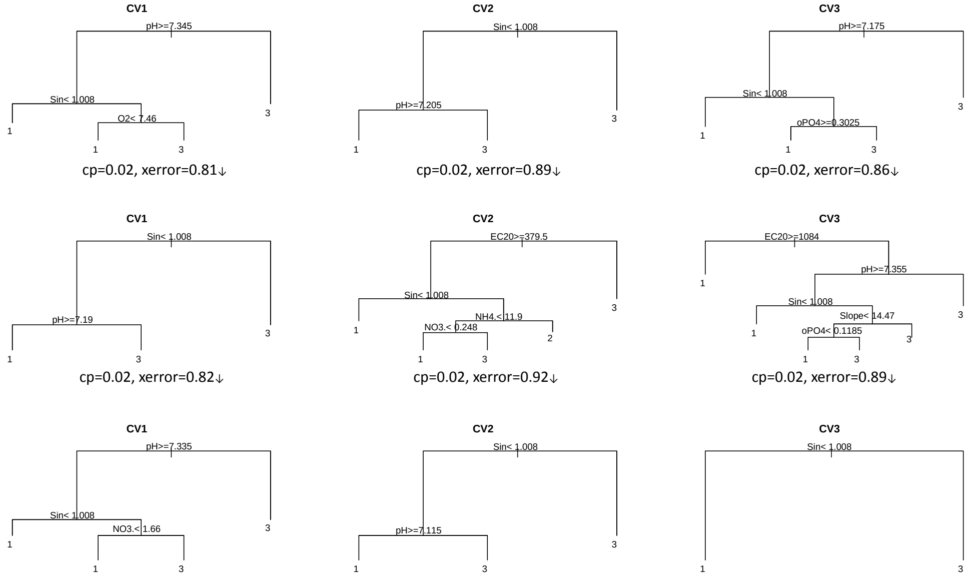
Figure S2 The 288 trees built for the two scenarios using the two types of dataset (field and combined data) and the two resampling approaches (random and stratified). The following slides, from slide number 1 to 20 and from 21 to 40, are the information and trees built using the field and combined data, respectively. For the hindcasting scenario, the models were calibrated using the data period D4 and were validated using the environmental predictor of the data period D4, D3, D2 and D1. For the forecasting scenario, the models were calibrated using the data period D1 and were validated using the environmental predictor of the data period D1, D2, D3 and D4. The calibrated models' complex parameter and their cross-validation error (xerror) are given under each tree. The increased xerror (↑) is highlighted in yellow. The decreased and stable xerrors are respectively labeled with (↓)and (~). The detailed information of each slide title is provided in the following:

- Field_Hind_V-D4(3,2,1)_Random/Stratified: models were calibrated using the field data, for hindcasting scenario (validation based on D4, D3, D2, D1), and using the random/stratified resampling approach.
- Field_Fore_V-D1(2,3,4)_Random/Stratified: models were calibrated using the filed data, for forecasting scenario (validation based on D1, D2, D3, D4), and using the random/stratified resampling approach.
- Comined_Hind_V-D4(3,2,1)_Random/Stratified: models were calibrated using the combined data, for hindcasting scenario (validation based on D4, D3, D2, D1), and using the random/ stratified resampling approach.
- Comined_Fore_V-D1(2,3,4)_Random/Stratified: models were calibrated using the combined data, for forecasting scenario (validation based on D1, D2, D3, D4), and using the random/stratified resampling approach.

Field data_Hindcasting_Random

Models calibrated on D4
and validated by D4,D3,D2,D1

Field_Hind_V-D4_Random



Replication 1

cp=0.02, xerror=0.81↓

cp=0.02, xerror=0.89↓

cp=0.02, xerror=0.86↓

Replication 2

cp=0.02, xerror=0.82↓

cp=0.02, xerror=0.92↓

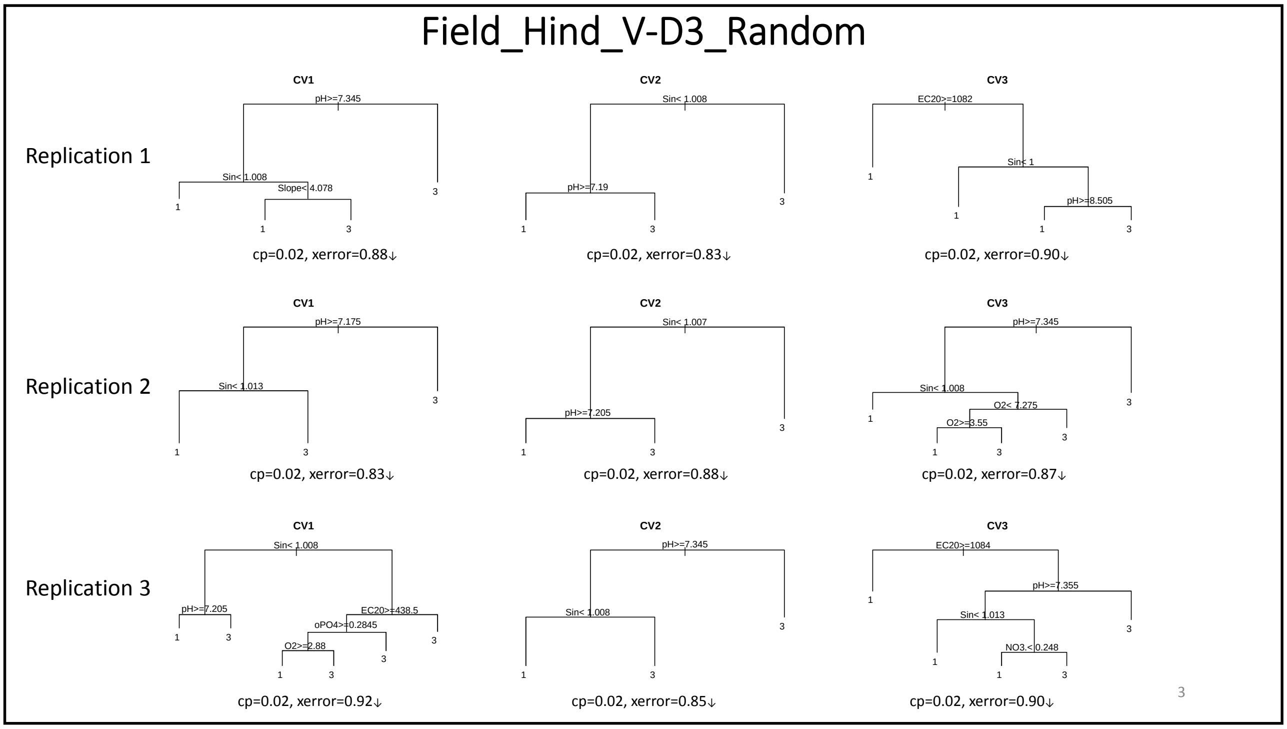
cp=0.02, xerror=0.89↓

Replication 3

cp=0.02, xerror=0.86↓

cp=0.02, xerror=0.87↓

cp=0.02, xerror=0.87↓



Field_Hind_V-D3_Random

Replication 1

cp=0.02, xerror=0.88↓ cp=0.02, xerror=0.83↓ cp=0.02, xerror=0.90↓

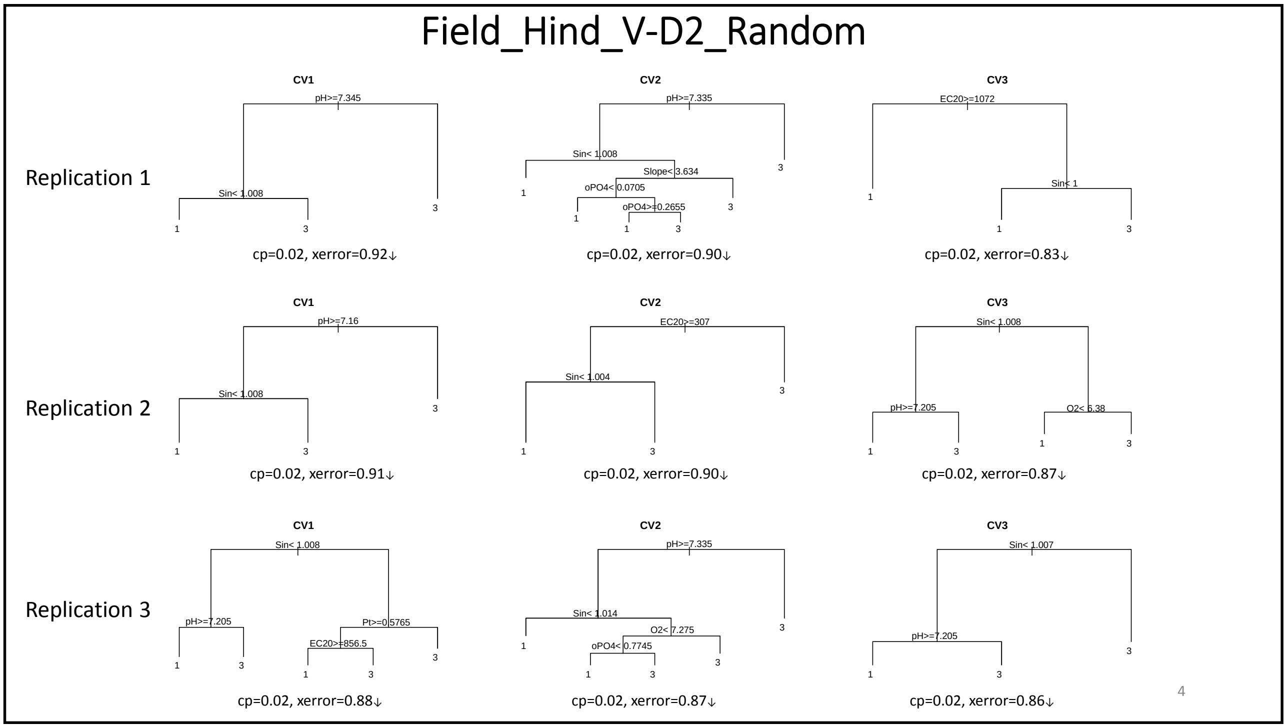
Replication 2

cp=0.02, xerror=0.83↓ cp=0.02, xerror=0.88↓ cp=0.02, xerror=0.87↓

Replication 3

cp=0.02, xerror=0.92↓ cp=0.02, xerror=0.85↓ cp=0.02, xerror=0.90↓

3



Field_Hind_V-D2_Random

Replication 1

cp=0.02, xerror=0.92↓ cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.83↓

Replication 2

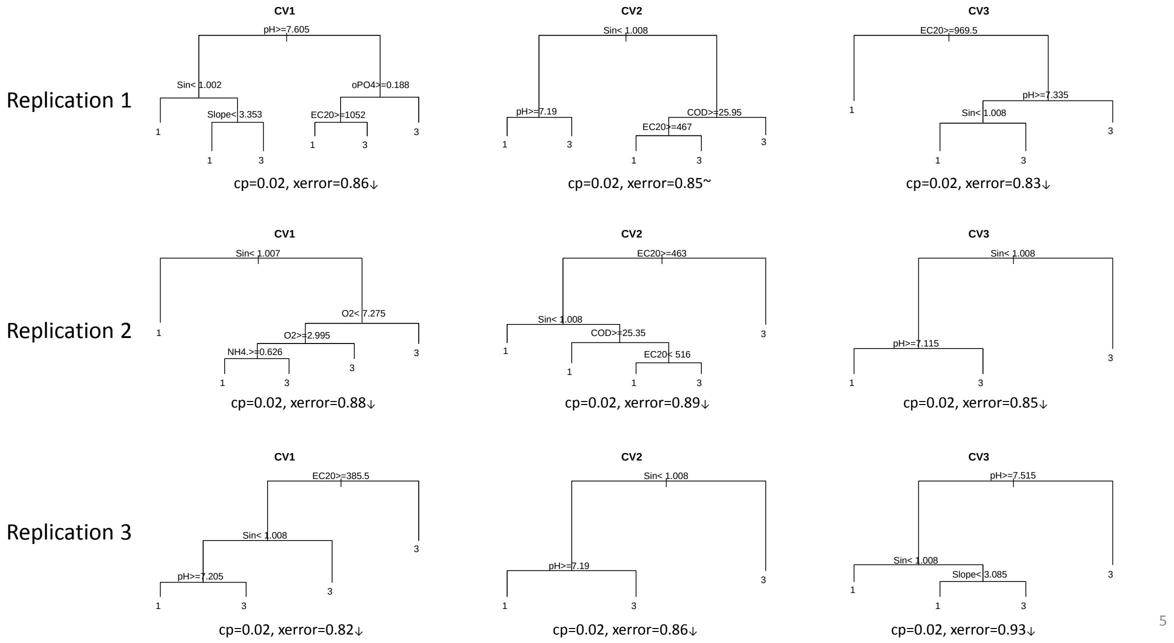
cp=0.02, xerror=0.91↓ cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.87↓

Replication 3

cp=0.02, xerror=0.88↓ cp=0.02, xerror=0.87↓ cp=0.02, xerror=0.86↓

4

Field_Hind_V-D1_Random



Replication 1

cp=0.02, xerror=0.86↓

cp=0.02, xerror=0.85~

cp=0.02, xerror=0.83↓

Replication 2

cp=0.02, xerror=0.88↓

cp=0.02, xerror=0.89↓

cp=0.02, xerror=0.85↓

Replication 3

cp=0.02, xerror=0.82↓

cp=0.02, xerror=0.86↓

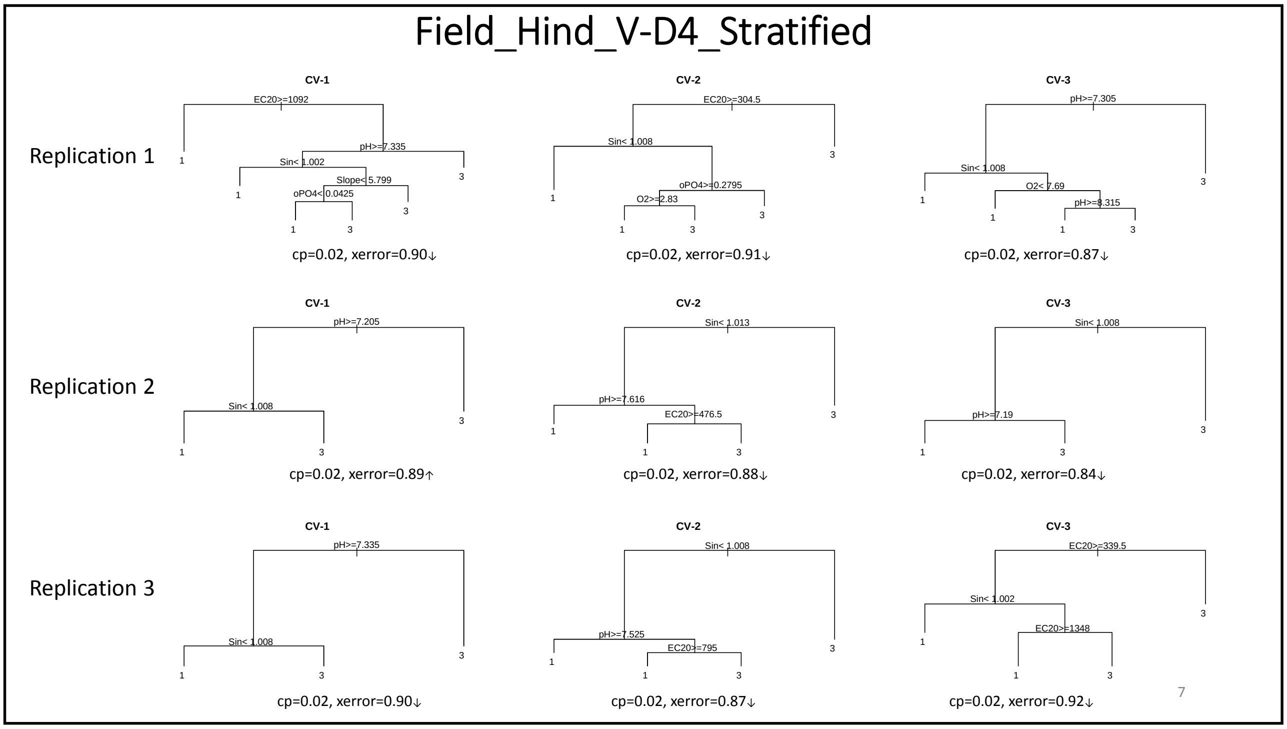
cp=0.02, xerror=0.93↓

5

Field data_Hindcasting_Stratified

Models calibrated on D4
and validated by D4,D3,D2,D1

6



Field_Hind_V-D4_Stratified

Replication 1

cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.91↓ cp=0.02, xerror=0.87↓

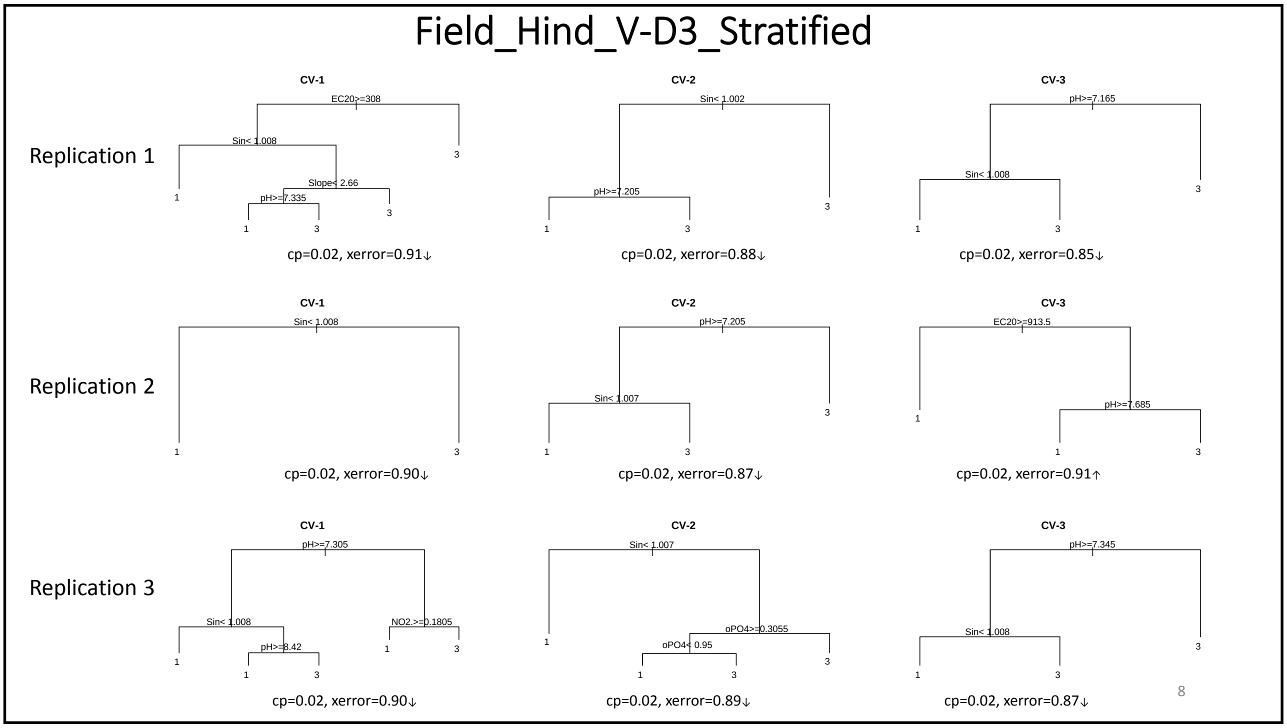
Replication 2

cp=0.02, xerror=0.89↑ cp=0.02, xerror=0.88↓ cp=0.02, xerror=0.84↓

Replication 3

cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.87↓ cp=0.02, xerror=0.92↓

7



Field_Hind_V-D3_Stratified

Replication 1

cp=0.02, xerror=0.91↓ cp=0.02, xerror=0.88↓ cp=0.02, xerror=0.85↓

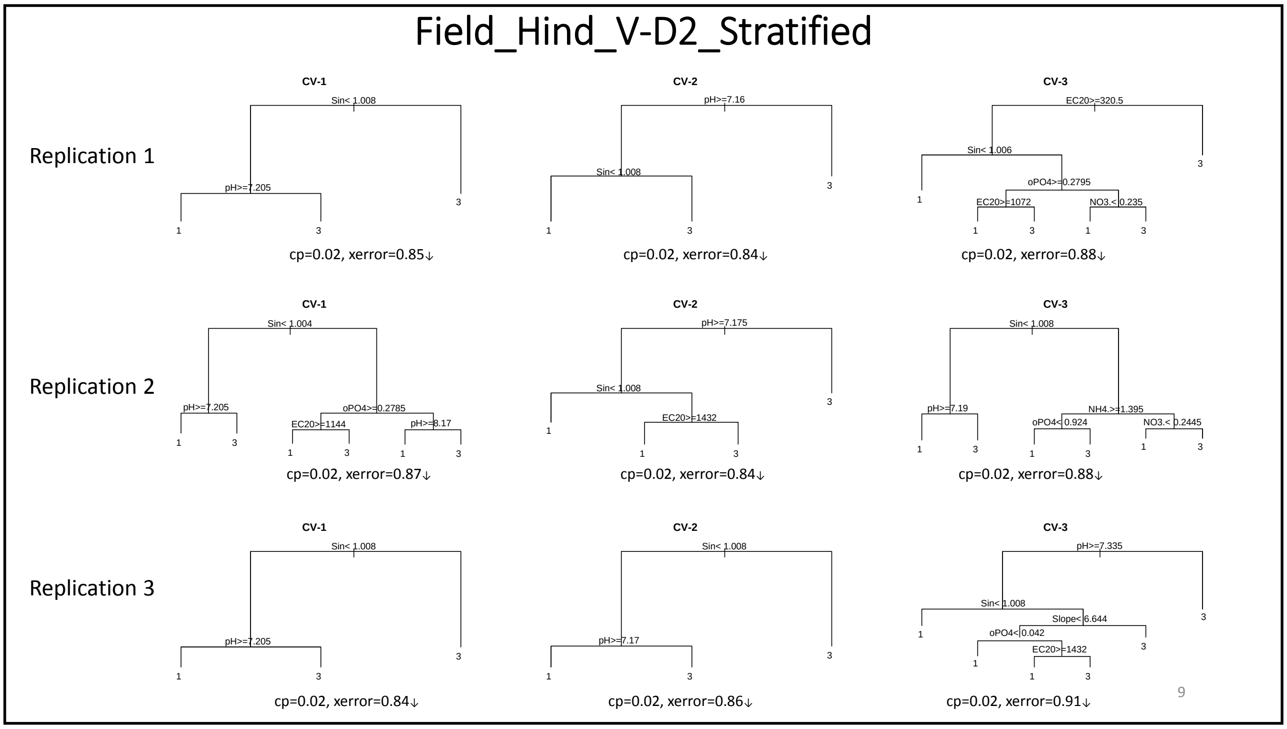
Replication 2

cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.87↓ cp=0.02, xerror=0.91↑

Replication 3

cp=0.02, xerror=0.90↓ cp=0.02, xerror=0.89↓ cp=0.02, xerror=0.87↓

8



Field_Hind_V-D2_Stratified

Replication 1

cp=0.02, xerror=0.85↓ cp=0.02, xerror=0.84↓ cp=0.02, xerror=0.88↓

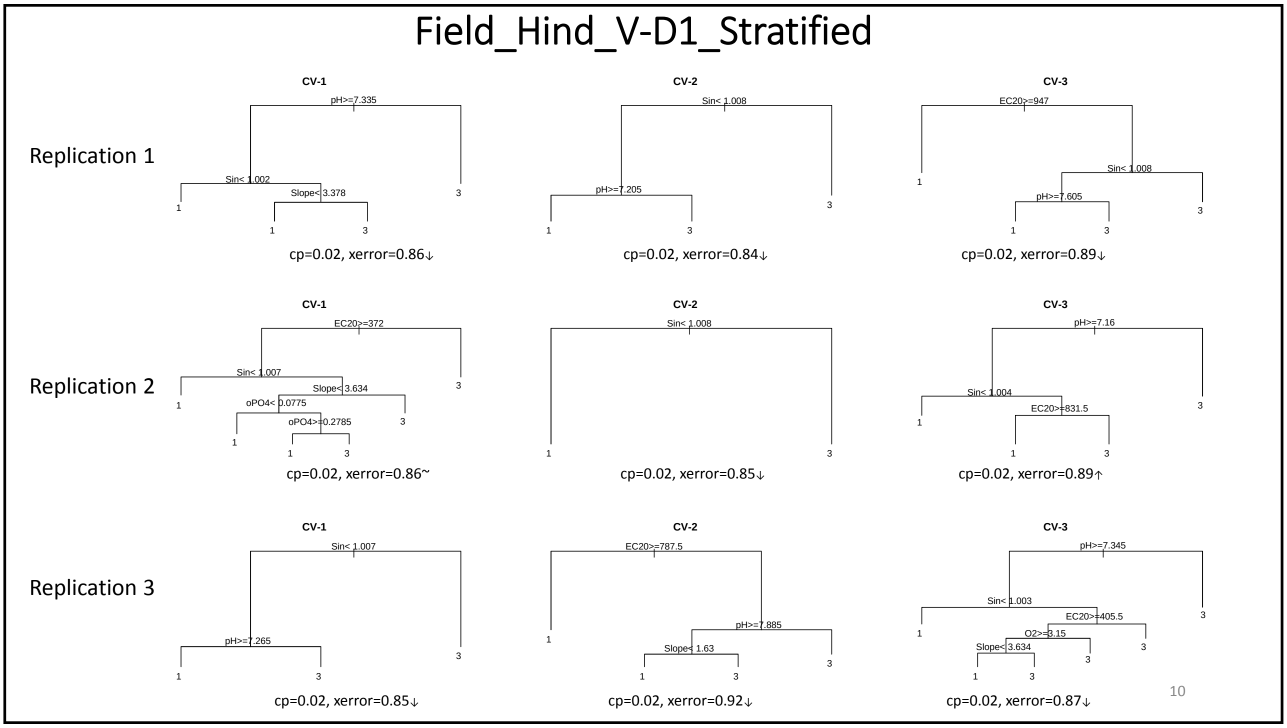
Replication 2

cp=0.02, xerror=0.87↓ cp=0.02, xerror=0.84↓ cp=0.02, xerror=0.88↓

Replication 3

cp=0.02, xerror=0.84↓ cp=0.02, xerror=0.86↓ cp=0.02, xerror=0.91↓

9



Field_Hind_V-D1_Stratified

Replication 1

cp=0.02, xerror=0.86↓ cp=0.02, xerror=0.84↓ cp=0.02, xerror=0.89↓

Replication 2

cp=0.02, xerror=0.86~ cp=0.02, xerror=0.85↓ cp=0.02, xerror=0.89↑

Replication 3

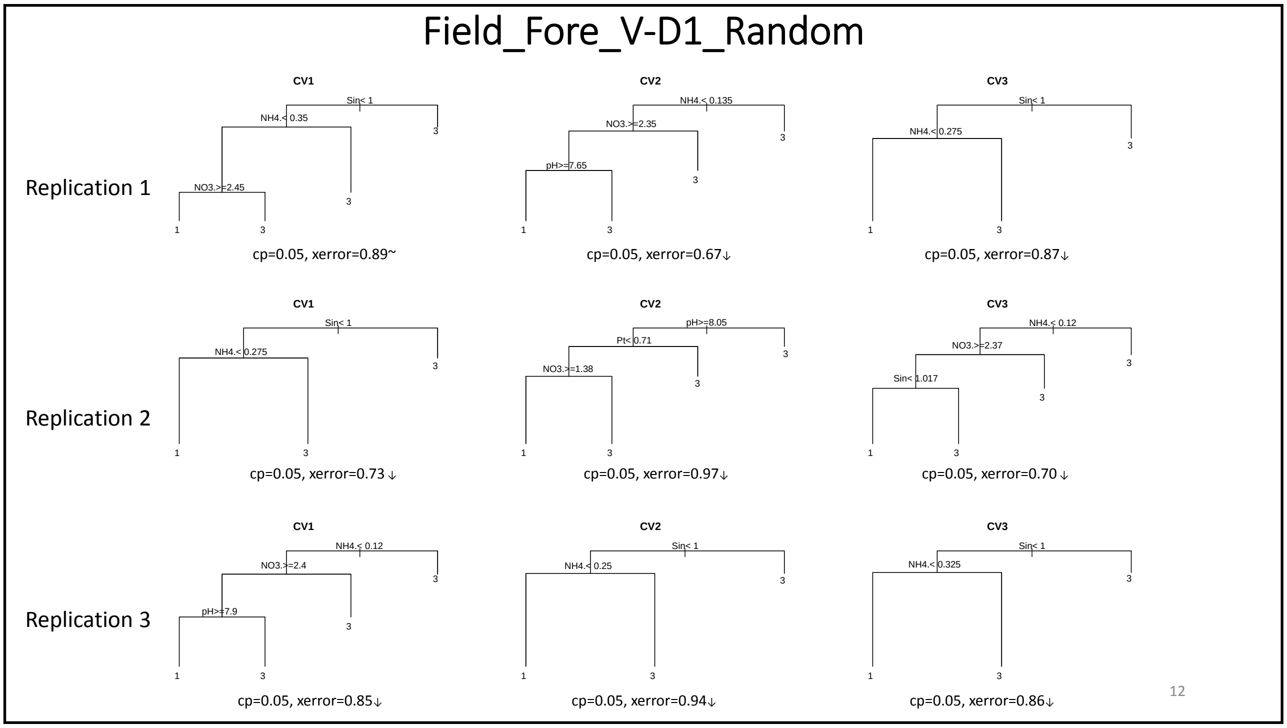
cp=0.02, xerror=0.85↓ cp=0.02, xerror=0.92↓ cp=0.02, xerror=0.87↓

10

Field data_Forecasting_Random

Models calibrated on D1
and validated by D1,D2,D3,D4

11



Field_Fore_V-D1_Random

Replication 1

cp=0.05, xerror=0.89~

cp=0.05, xerror=0.67↓

cp=0.05, xerror=0.87↓

Replication 2

cp=0.05, xerror=0.73↓

cp=0.05, xerror=0.97↓

cp=0.05, xerror=0.70↓

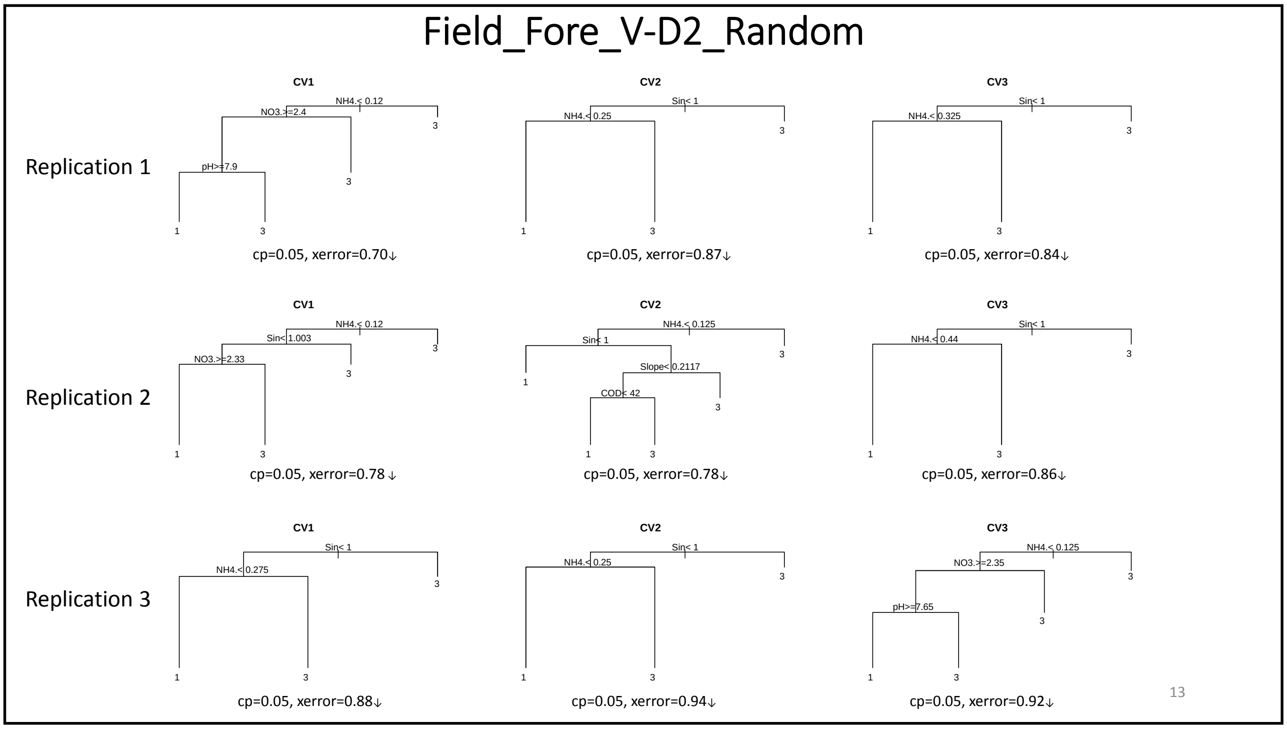
Replication 3

cp=0.05, xerror=0.85↓

cp=0.05, xerror=0.94↓

cp=0.05, xerror=0.86↓

12



Field_Fore_V-D2_Random

Replication 1

cp=0.05, xerror=0.70↓ cp=0.05, xerror=0.87↓ cp=0.05, xerror=0.84↓

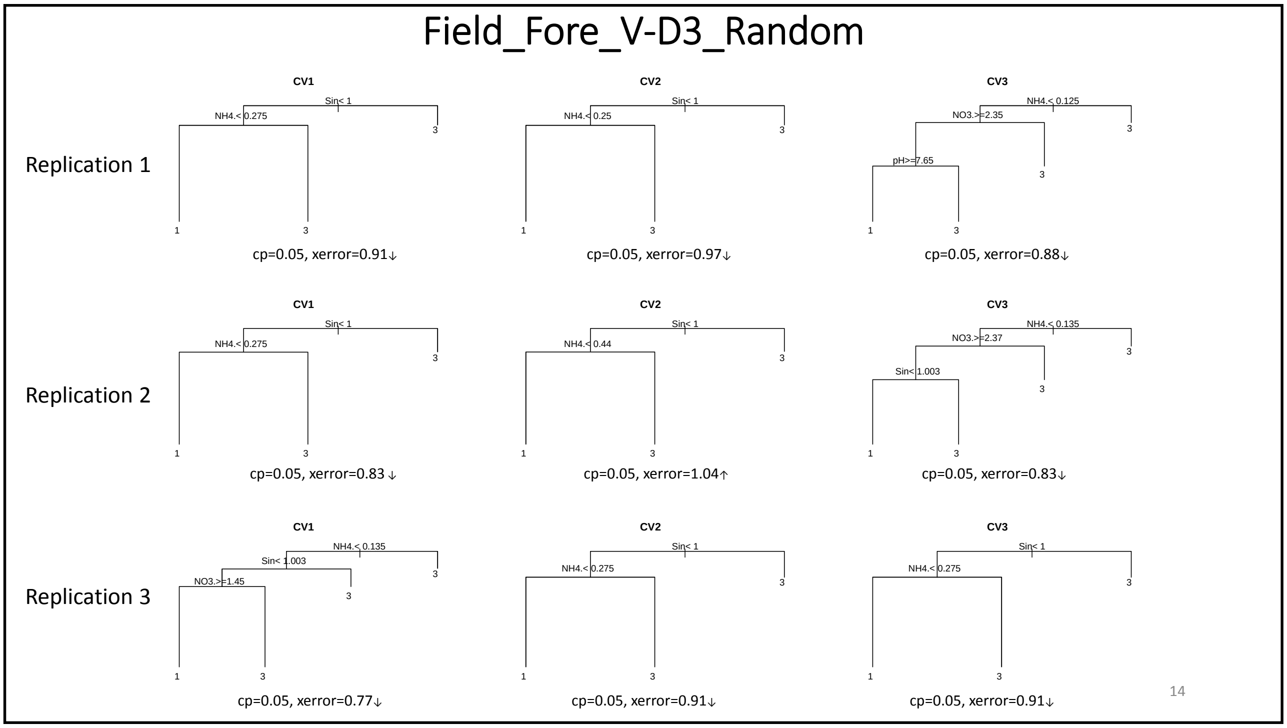
Replication 2

cp=0.05, xerror=0.78↓ cp=0.05, xerror=0.78↓ cp=0.05, xerror=0.86↓

Replication 3

cp=0.05, xerror=0.88↓ cp=0.05, xerror=0.94↓ cp=0.05, xerror=0.92↓

13



Field_Fore_V-D3_Random

Replication 1

cp=0.05, xerror=0.91↓ cp=0.05, xerror=0.97↓ cp=0.05, xerror=0.88↓

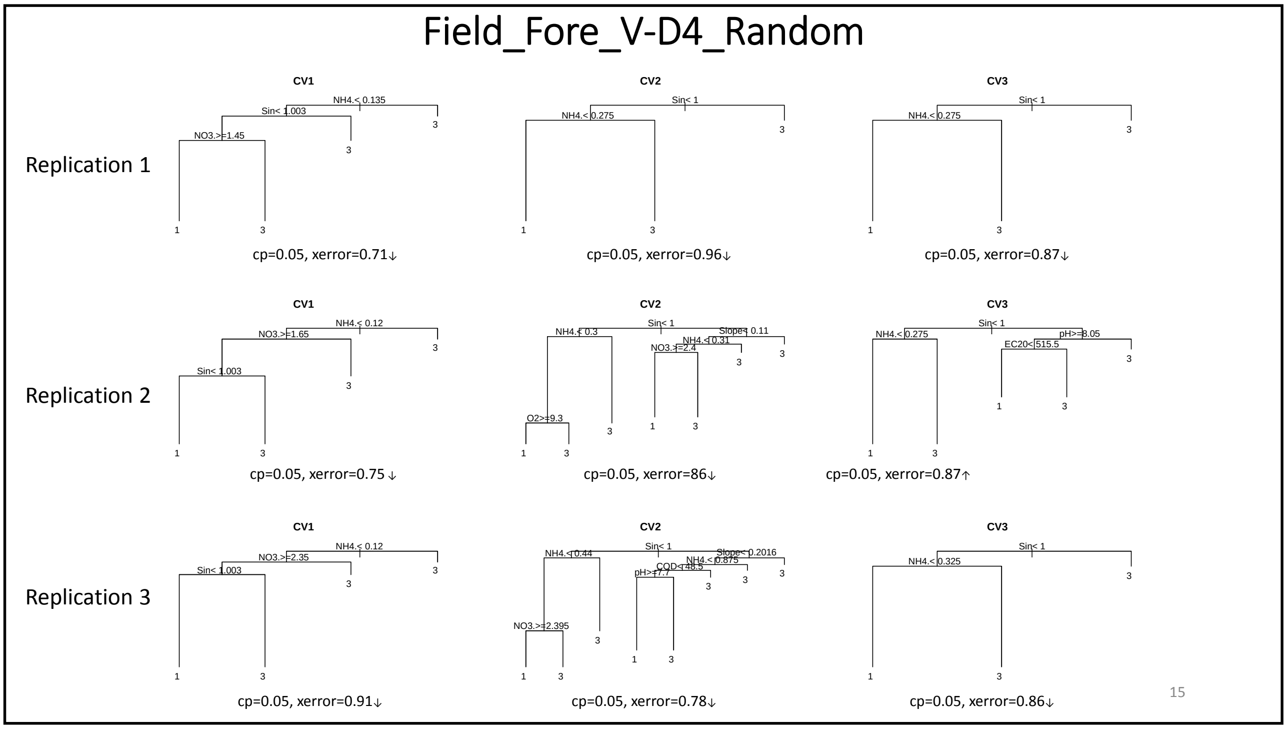
Replication 2

cp=0.05, xerror=0.83↓ cp=0.05, xerror=1.04↑ cp=0.05, xerror=0.83↓

Replication 3

cp=0.05, xerror=0.77↓ cp=0.05, xerror=0.91↓ cp=0.05, xerror=0.91↓

14



Field_Fore_V-D4_Random

Replication 1

cp=0.05, xerror=0.71↓

cp=0.05, xerror=0.96↓

cp=0.05, xerror=0.87↓

Replication 2

cp=0.05, xerror=0.75 ↓

cp=0.05, xerror=86↓

cp=0.05, xerror=0.87↑

Replication 3

cp=0.05, xerror=0.91↓

cp=0.05, xerror=0.78↓

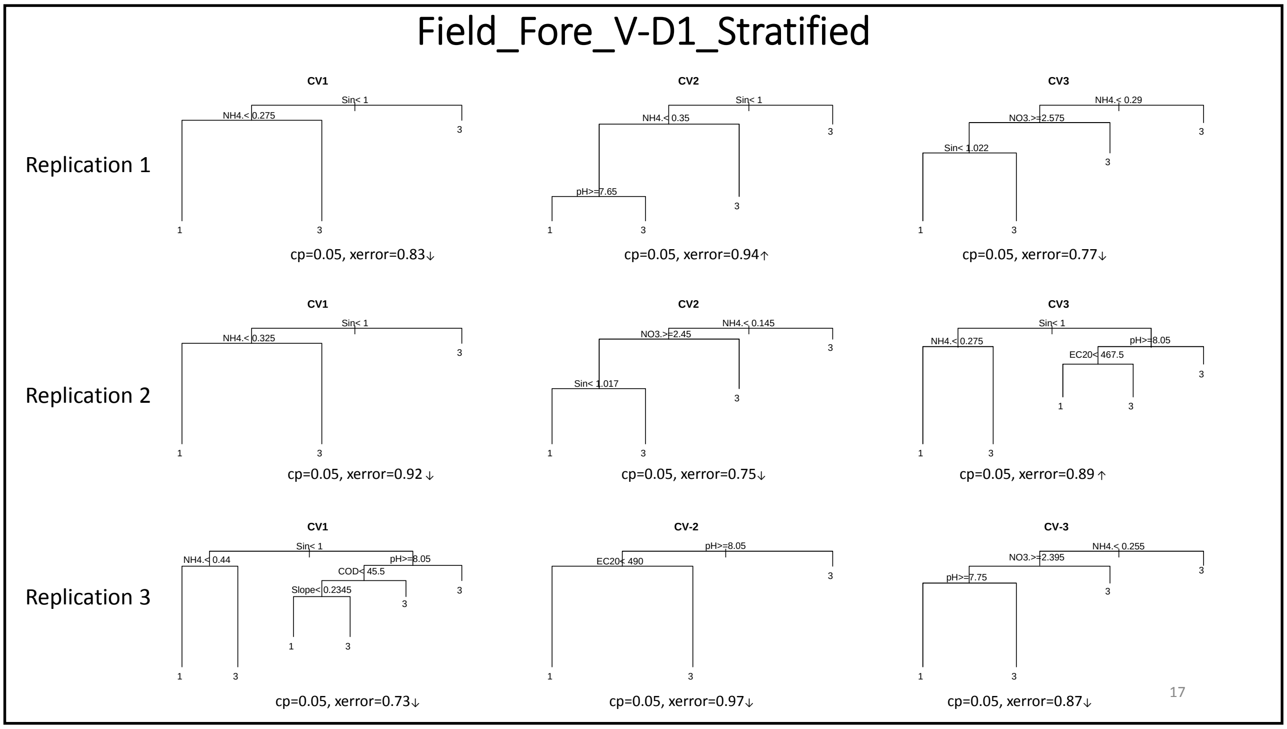
cp=0.05, xerror=0.86↓

15

Field data_Forecasting_Stratified

Models calibrated on D1
and validated by D1,D2,D3,D4

16



Field_Fore_V-D1_Stratified

Replication 1

cp=0.05, xerror=0.83↓

cp=0.05, xerror=0.94↑

cp=0.05, xerror=0.77↓

Replication 2

cp=0.05, xerror=0.92↓

cp=0.05, xerror=0.75↓

cp=0.05, xerror=0.89↑

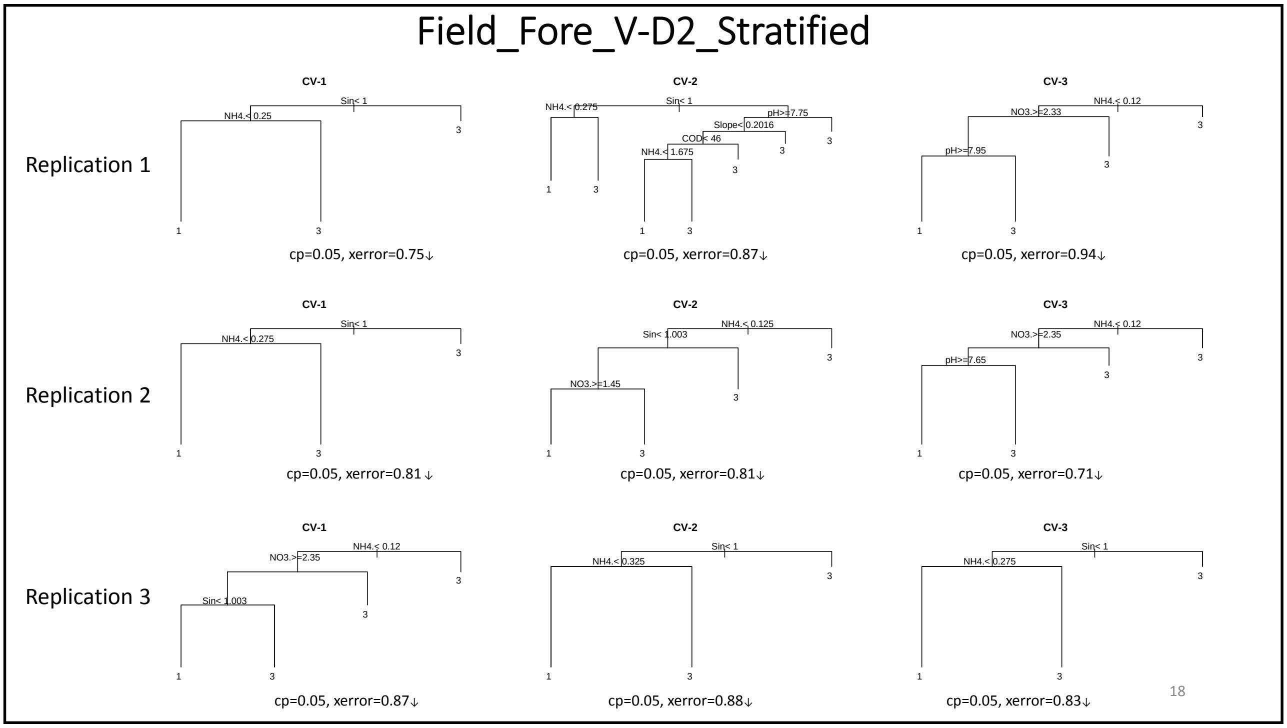
Replication 3

cp=0.05, xerror=0.73↓

cp=0.05, xerror=0.97↓

cp=0.05, xerror=0.87↓

17



Field_Fore_V-D2_Stratified

Replication 1

cp=0.05, xerror=0.75↓

cp=0.05, xerror=0.87↓

cp=0.05, xerror=0.94↓

Replication 2

cp=0.05, xerror=0.81↓

cp=0.05, xerror=0.81↓

cp=0.05, xerror=0.71↓

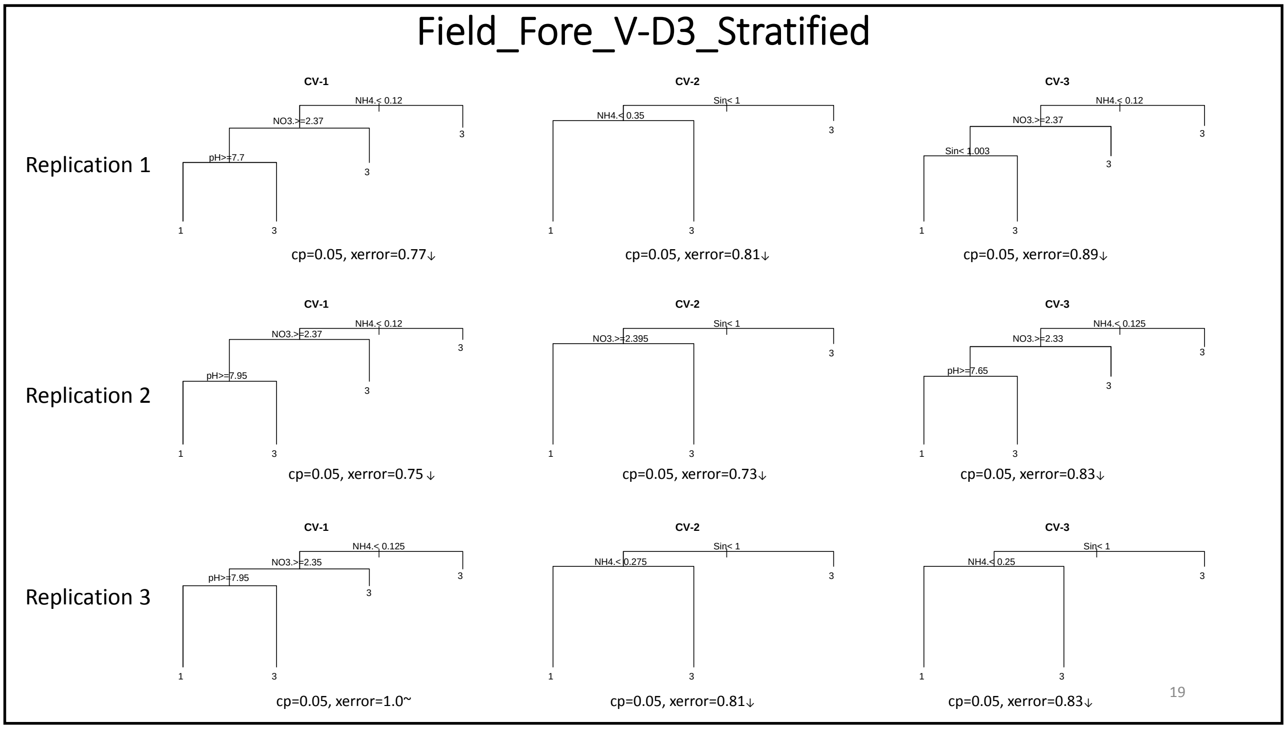
Replication 3

cp=0.05, xerror=0.87↓

cp=0.05, xerror=0.88↓

cp=0.05, xerror=0.83↓

18



Field_Fore_V-D3_Stratified

Replication 1

cp=0.05, xerror=0.77↓ cp=0.05, xerror=0.81↓ cp=0.05, xerror=0.89↓

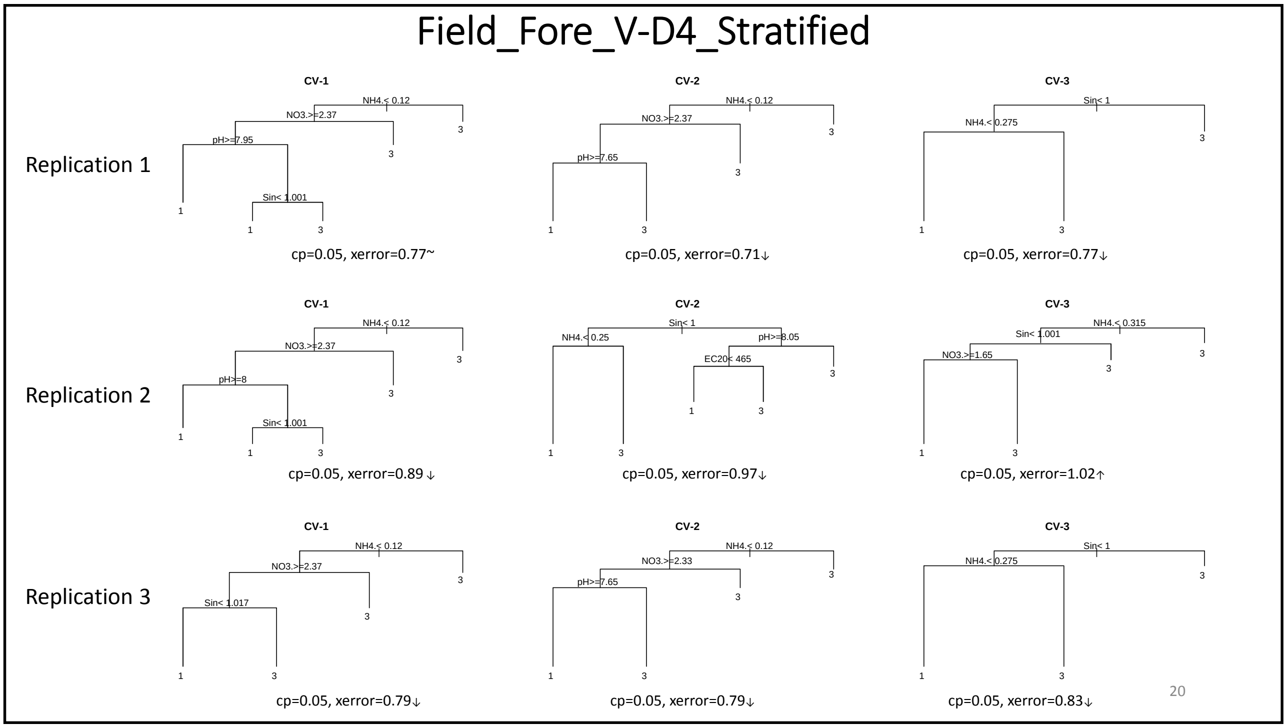
Replication 2

cp=0.05, xerror=0.75 ↓ cp=0.05, xerror=0.73↓ cp=0.05, xerror=0.83↓

Replication 3

cp=0.05, xerror=1.0~ cp=0.05, xerror=0.81↓ cp=0.05, xerror=0.83↓

19



Field_Fore_V-D4_Stratified

Replication 1

cp=0.05, xerror=0.77~ cp=0.05, xerror=0.71↓ cp=0.05, xerror=0.77↓

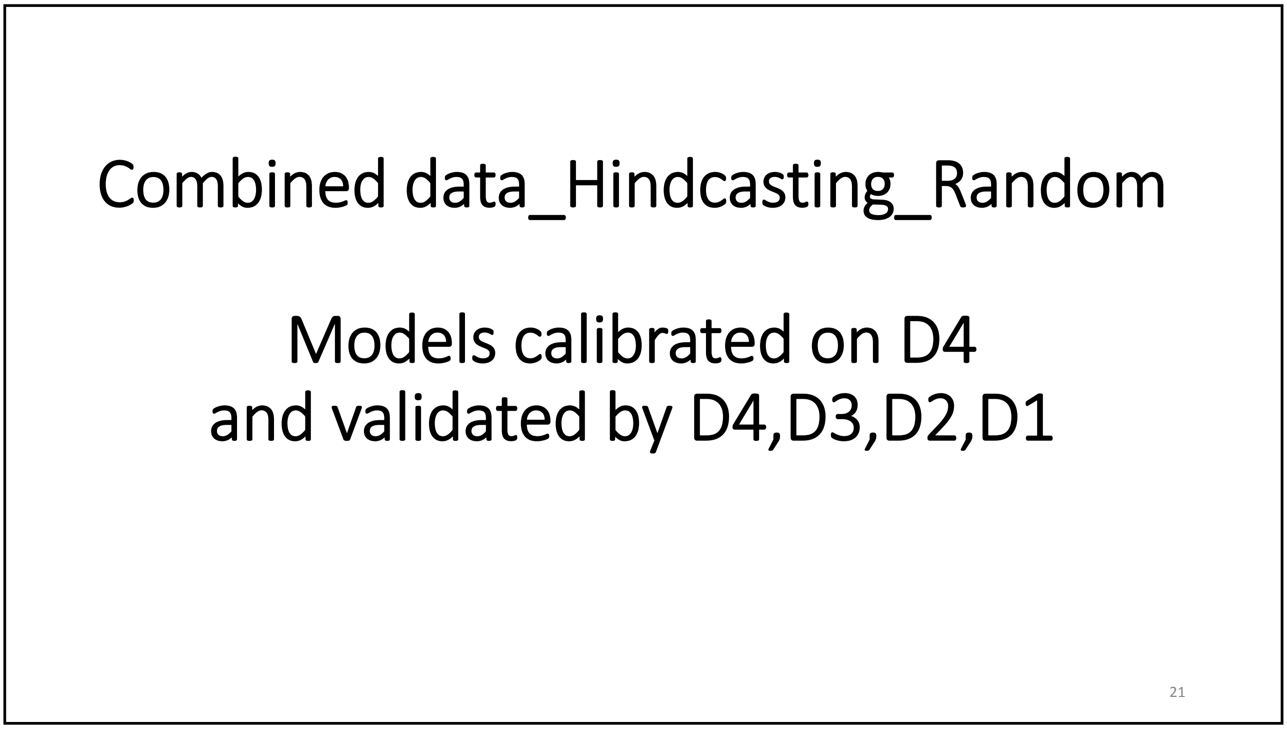
Replication 2

cp=0.05, xerror=0.89 ↓ cp=0.05, xerror=0.97↓ cp=0.05, xerror=1.02↑

Replication 3

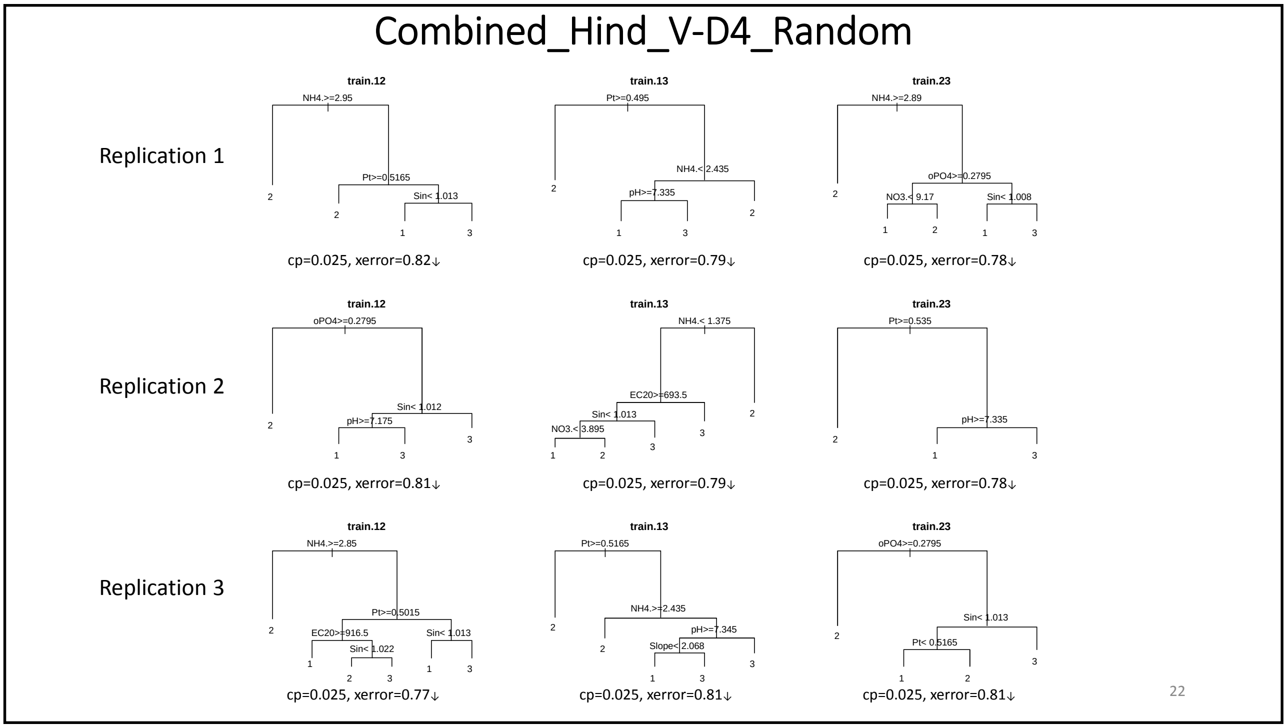
cp=0.05, xerror=0.79↓ cp=0.05, xerror=0.79↓ cp=0.05, xerror=0.83↓

20



Combined data_Hindcasting_Random

Models calibrated on D4
and validated by D4,D3,D2,D1



Combined_Hind_V-D4_Random

Replication 1

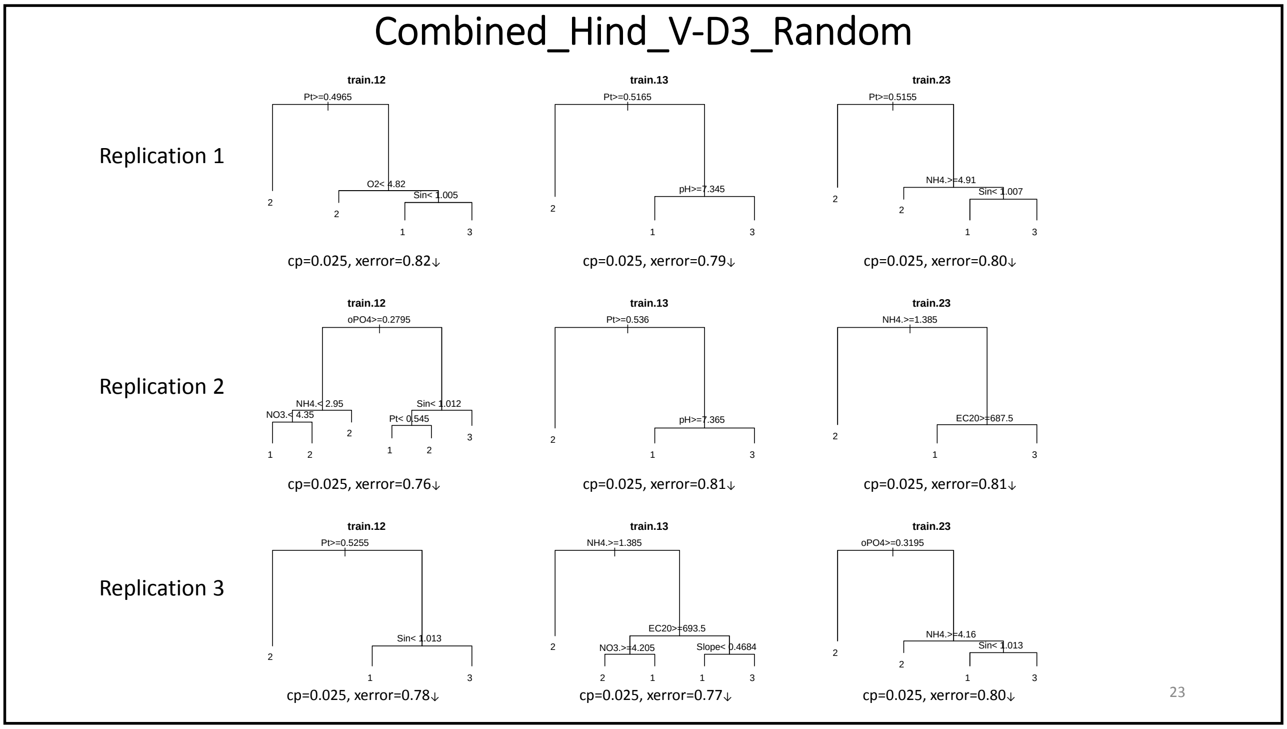
cp=0.025, xerror=0.82↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.78↓

Replication 2

cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.78↓

Replication 3

cp=0.025, xerror=0.77↓ cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.81↓



Combined_Hind_V-D3_Random

Replication 1

train.12

cp=0.025, xerror=0.82↓

train.13

cp=0.025, xerror=0.79↓

train.23

cp=0.025, xerror=0.80↓

Replication 2

train.12

cp=0.025, xerror=0.76↓

train.13

cp=0.025, xerror=0.81↓

train.23

cp=0.025, xerror=0.81↓

Replication 3

train.12

cp=0.025, xerror=0.78↓

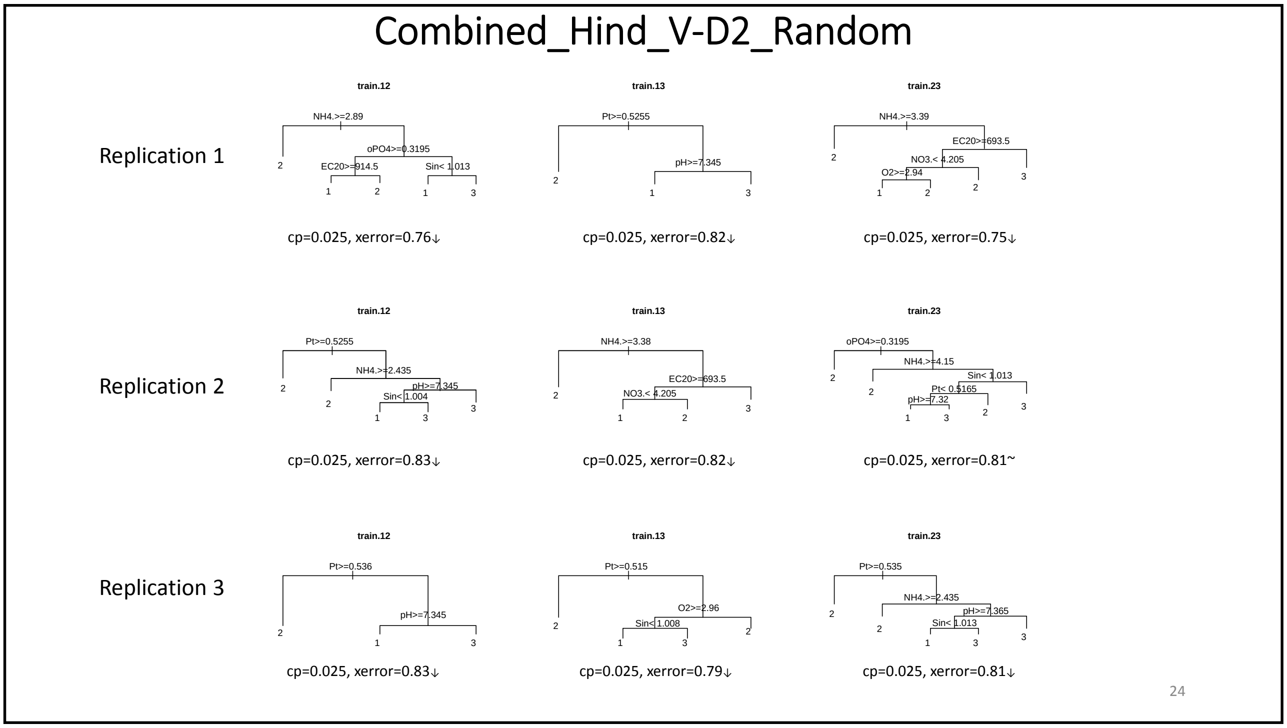
train.13

cp=0.025, xerror=0.77↓

train.23

cp=0.025, xerror=0.80↓

23



Combined_Hind_V-D2_Random

Replication 1

train.12

cp=0.025, xerror=0.76↓

train.13

cp=0.025, xerror=0.82↓

train.23

cp=0.025, xerror=0.75↓

Replication 2

train.12

cp=0.025, xerror=0.83↓

train.13

cp=0.025, xerror=0.82↓

train.23

cp=0.025, xerror=0.81~

Replication 3

train.12

cp=0.025, xerror=0.83↓

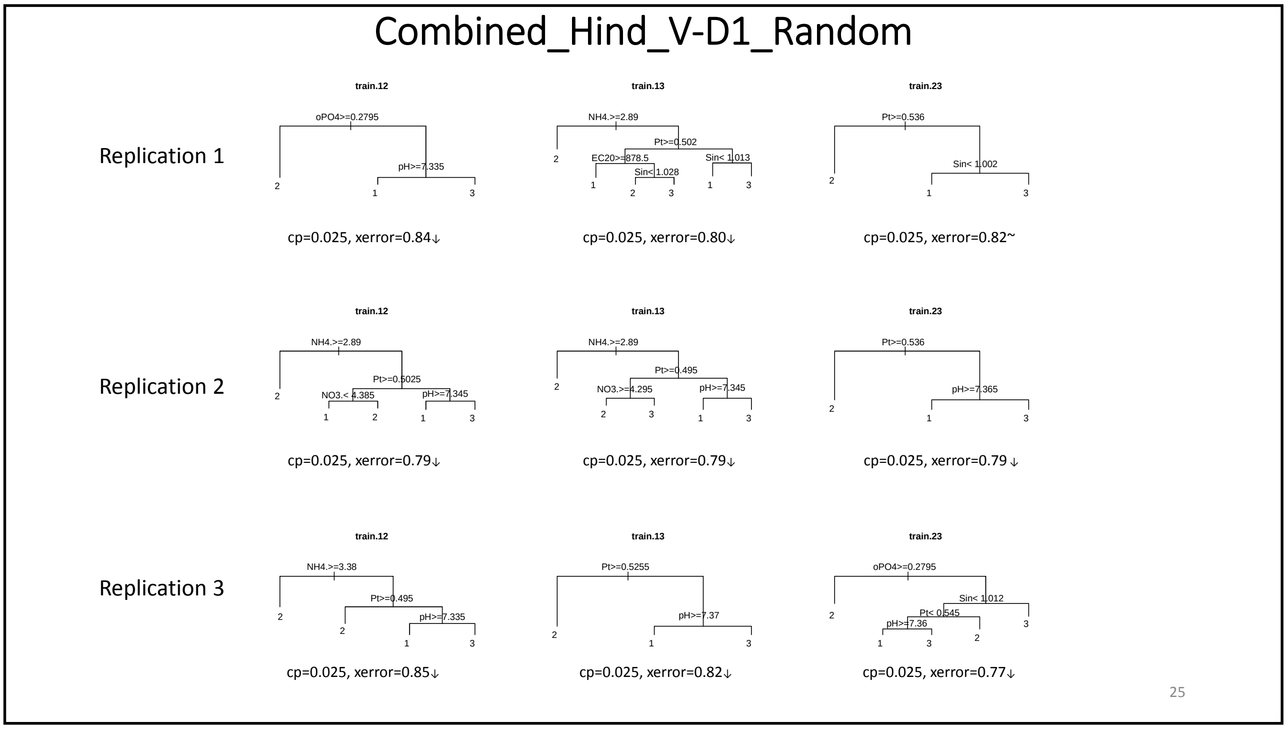
train.13

cp=0.025, xerror=0.79↓

train.23

cp=0.025, xerror=0.81↓

24



Combined_Hind_V-D1_Random

Replication 1

cp=0.025, xerror=0.84↓ cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.82~

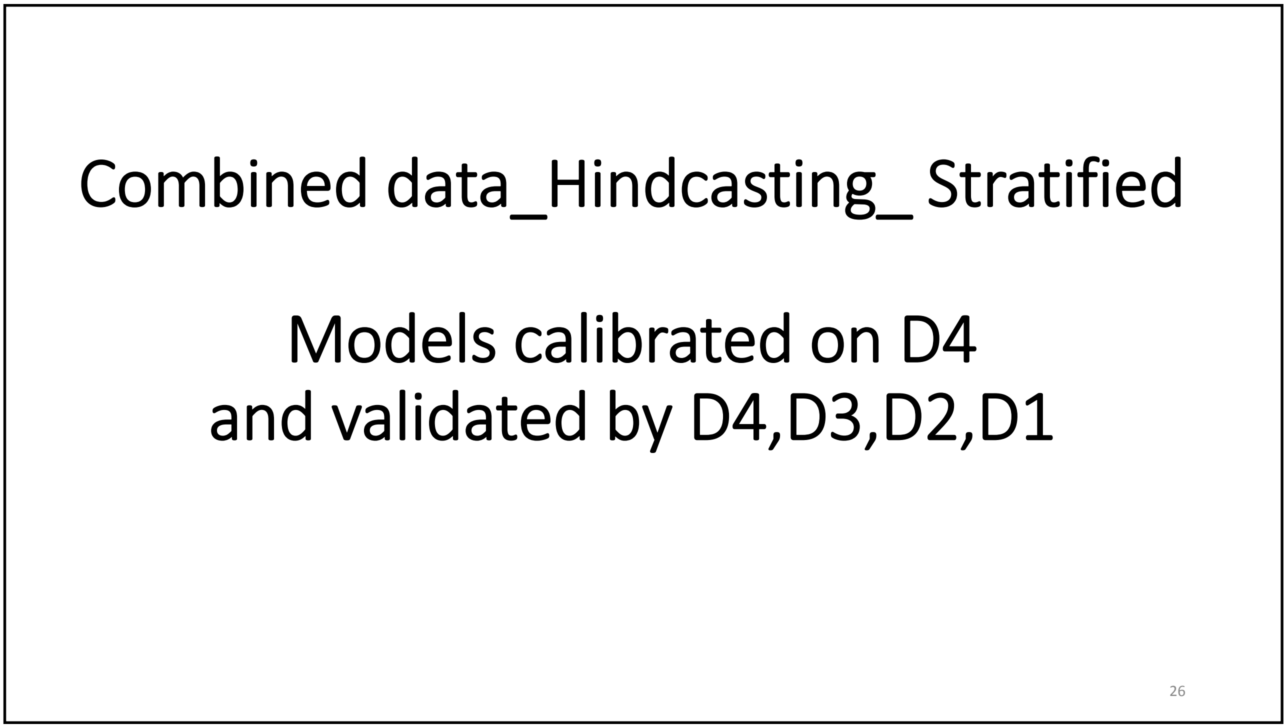
Replication 2

cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.79↓

Replication 3

cp=0.025, xerror=0.85↓ cp=0.025, xerror=0.82↓ cp=0.025, xerror=0.77↓

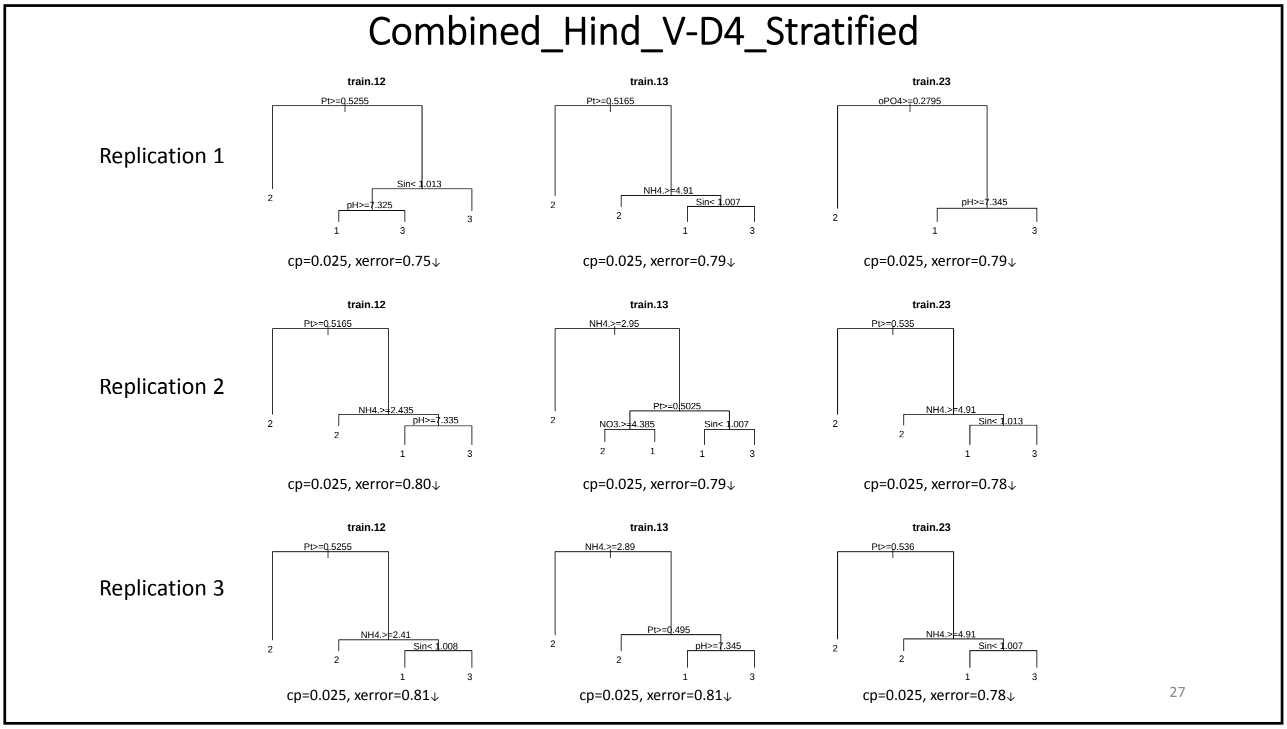
25



Combined data_Hindcasting_ Stratified

Models calibrated on D4
and validated by D4,D3,D2,D1

26



Combined_Hind_V-D4_Stratified

Replication 1

cp=0.025, xerror=0.75↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.79↓

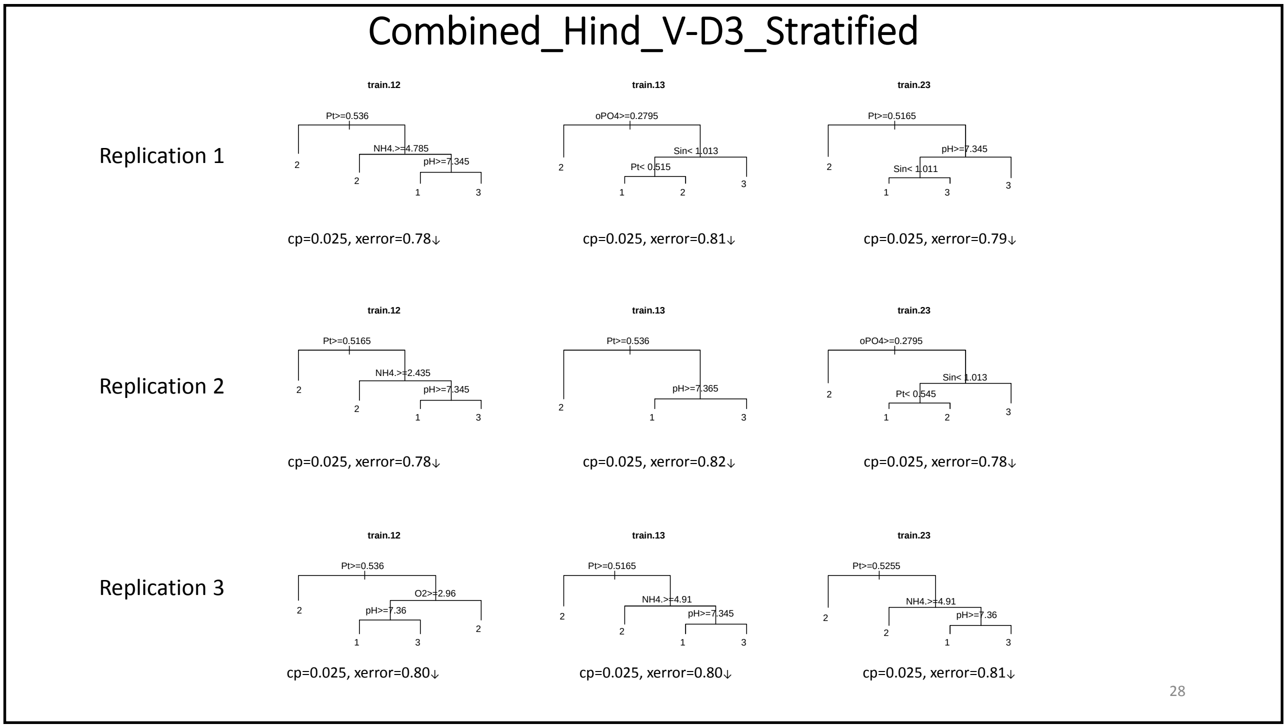
Replication 2

cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.78↓

Replication 3

cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.78↓

27



Combined_Hind_V-D3_Stratified

Replication 1

cp=0.025, xerror=0.78↓ cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.79↓

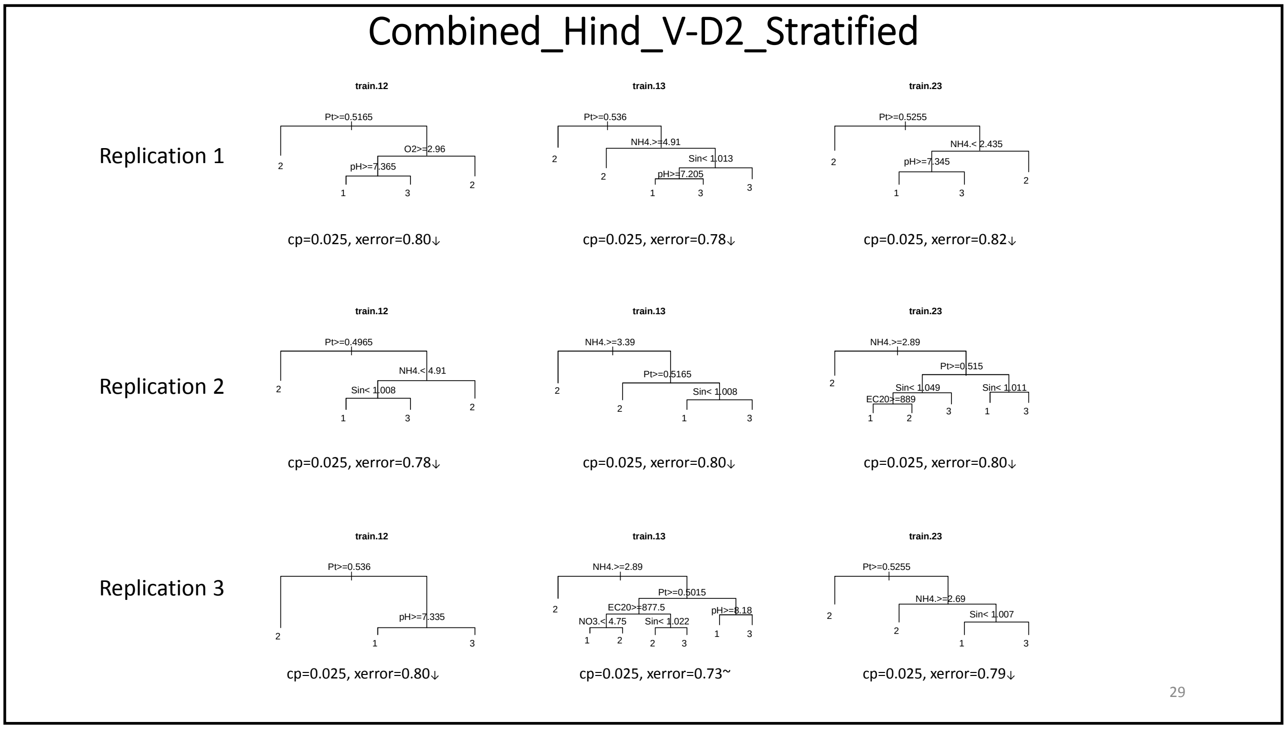
Replication 2

cp=0.025, xerror=0.78↓ cp=0.025, xerror=0.82↓ cp=0.025, xerror=0.78↓

Replication 3

cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.81↓

28



Combined_Hind_V-D2_Stratified

Replication 1

cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.78↓ cp=0.025, xerror=0.82↓

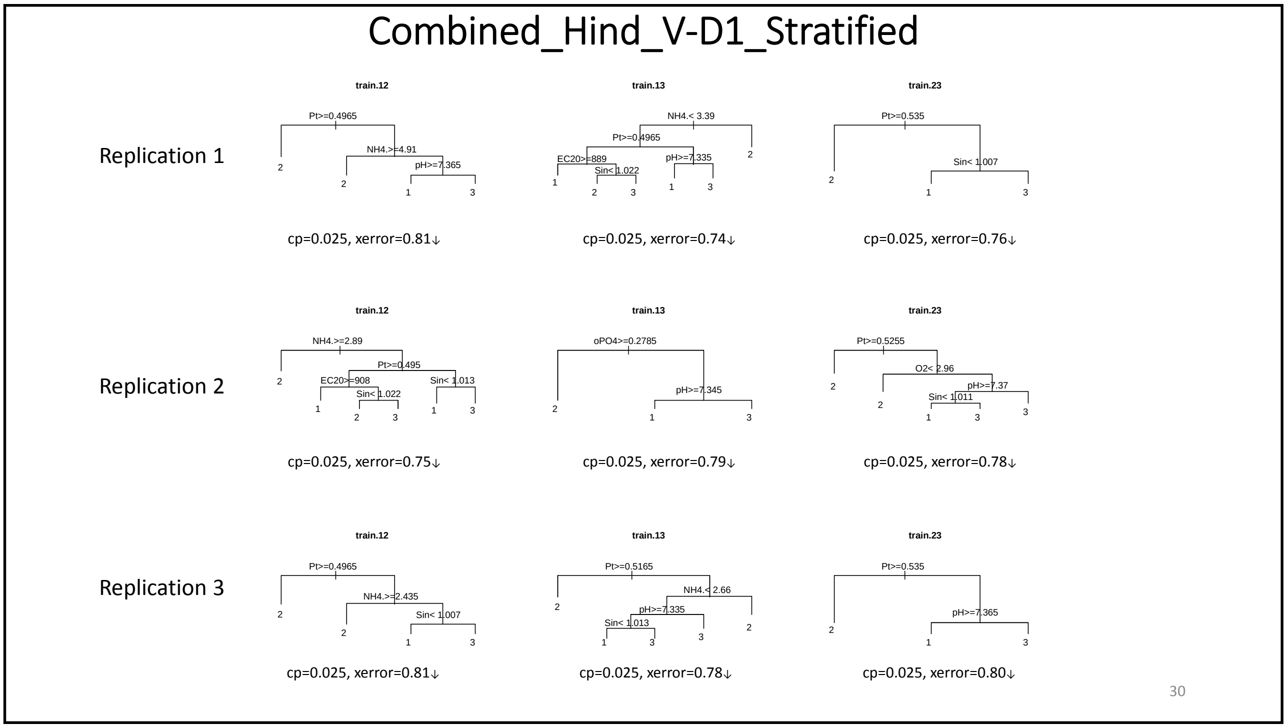
Replication 2

cp=0.025, xerror=0.78↓ cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.80↓

Replication 3

cp=0.025, xerror=0.80↓ cp=0.025, xerror=0.73~ cp=0.025, xerror=0.79↓

29



Combined_Hind_V-D1_Stratified

Replication 1

cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.74↓ cp=0.025, xerror=0.76↓

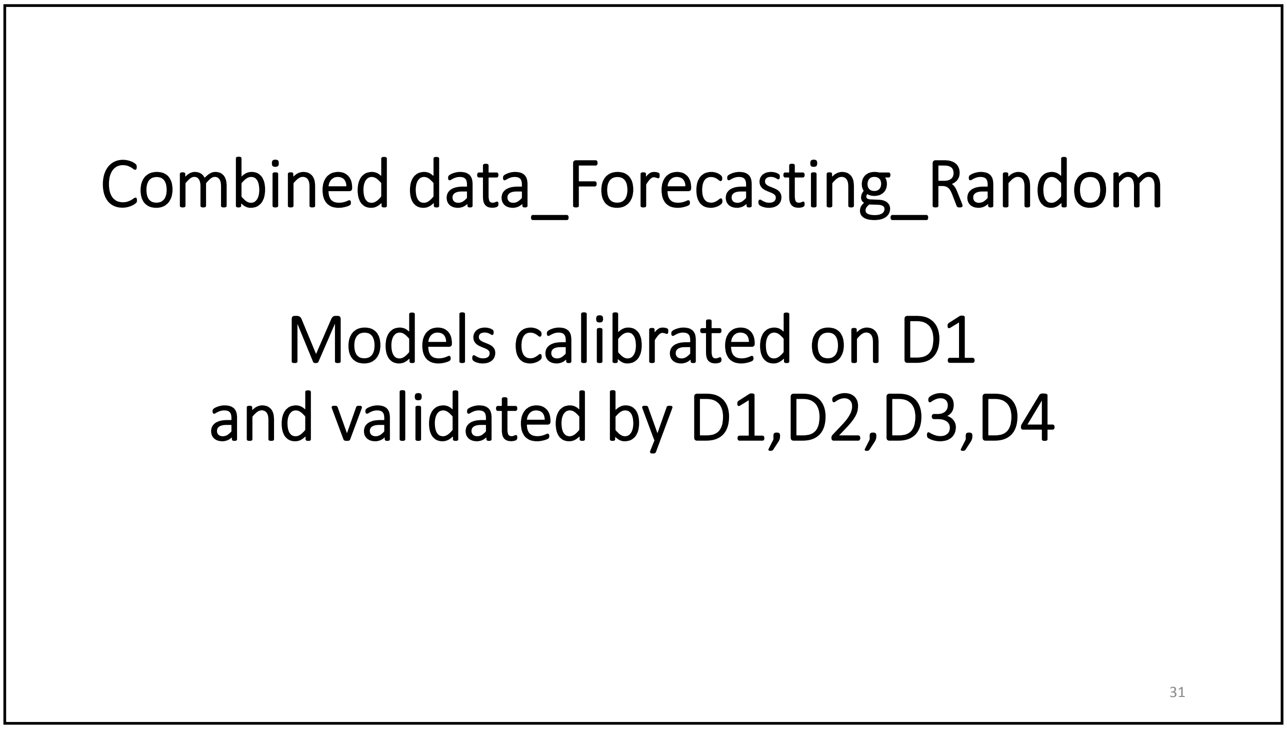
Replication 2

cp=0.025, xerror=0.75↓ cp=0.025, xerror=0.79↓ cp=0.025, xerror=0.78↓

Replication 3

cp=0.025, xerror=0.81↓ cp=0.025, xerror=0.78↓ cp=0.025, xerror=0.80↓

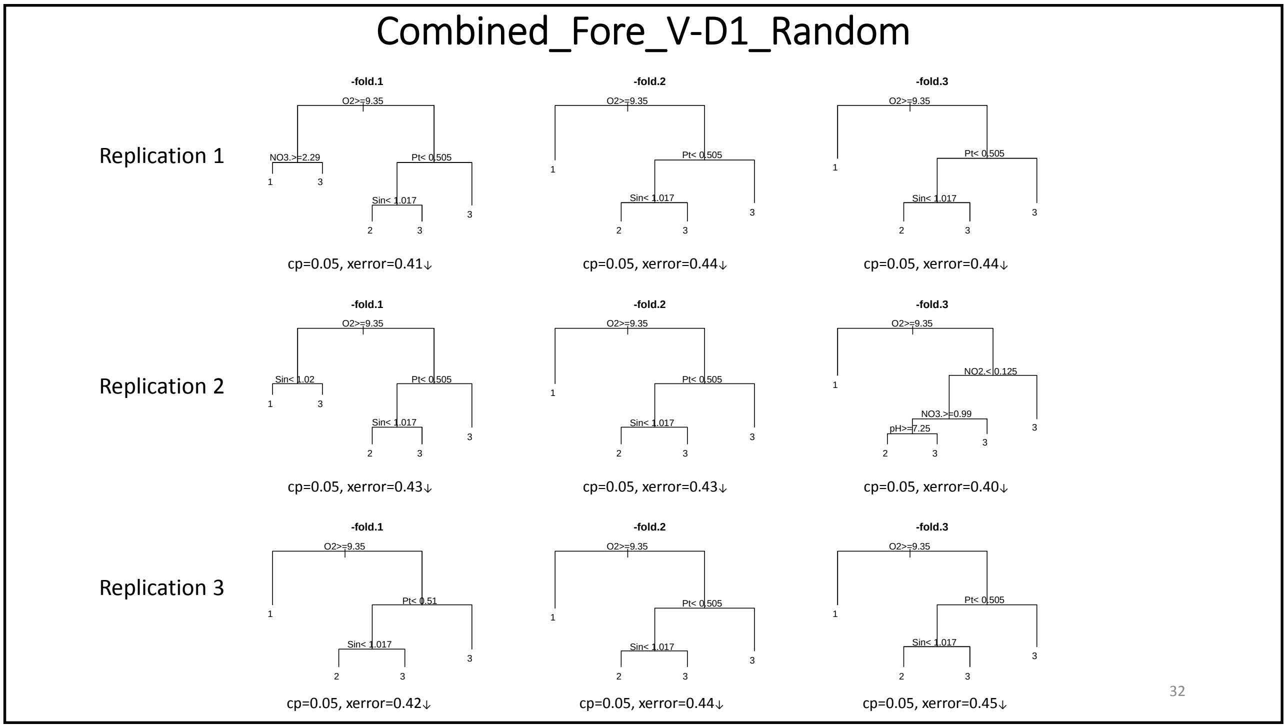
30



Combined data_Forecasting_Random

Models calibrated on D1 and validated by D1,D2,D3,D4

31



Combined_Fore_V-D1_Random

Replication 1

cp=0.05, xerror=0.41↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.44↓

Replication 2

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.40↓

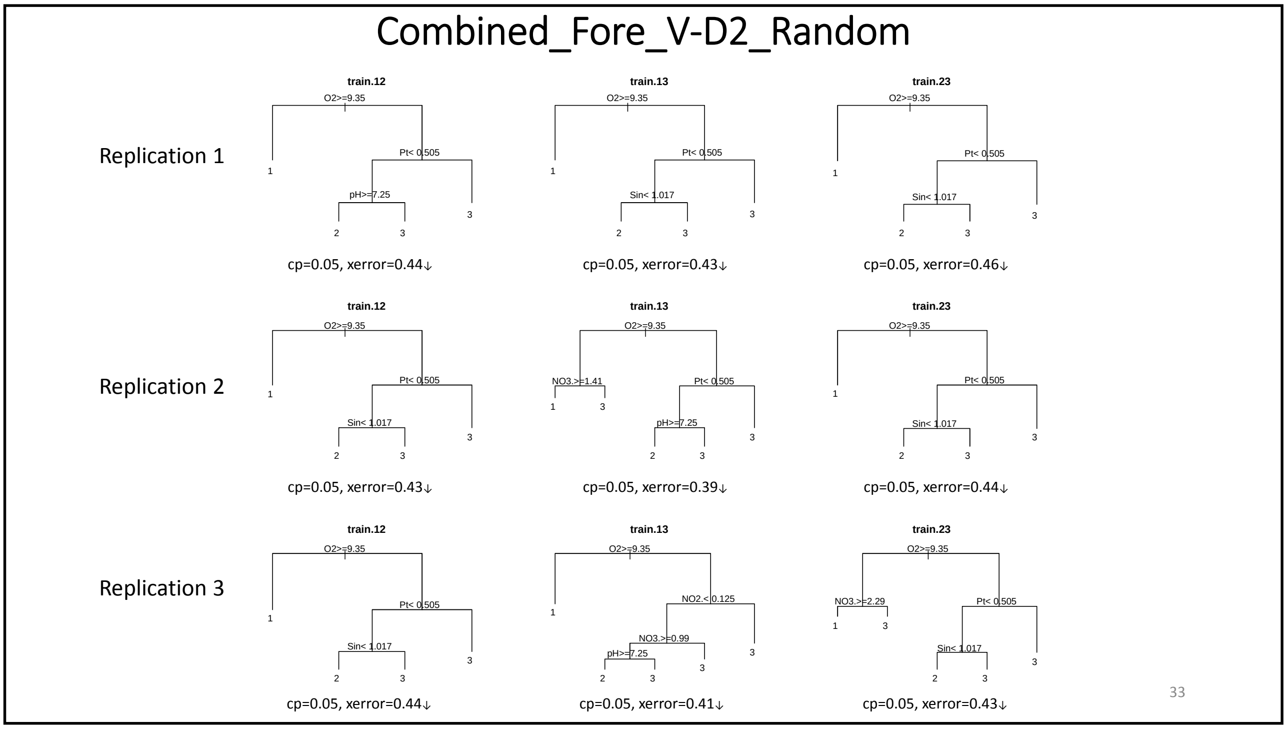
Replication 3

cp=0.05, xerror=0.42↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.45↓

32



Combined_Fore_V-D2_Random

Replication 1

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.46↓

Replication 2

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.39↓

cp=0.05, xerror=0.44↓

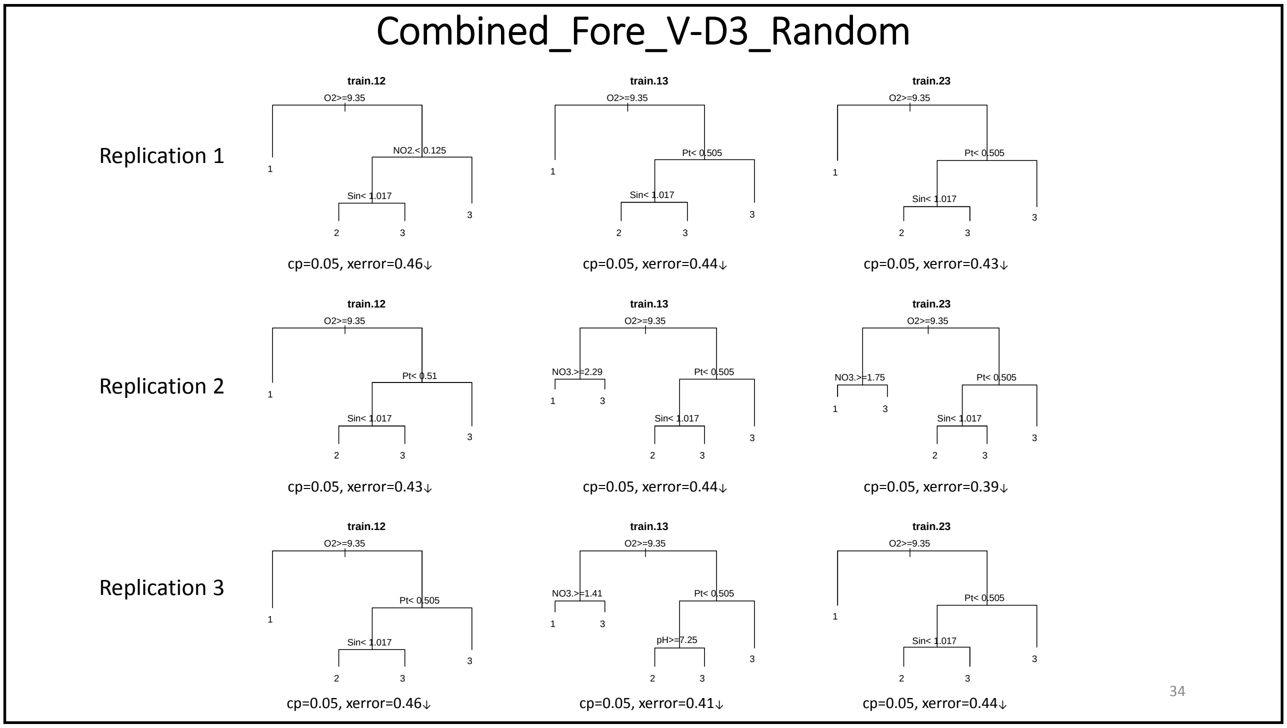
Replication 3

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.41↓

cp=0.05, xerror=0.43↓

33



Combined_Fore_V-D3_Random

Replication 1

cp=0.05, xerror=0.46↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.43↓

Replication 2

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.39↓

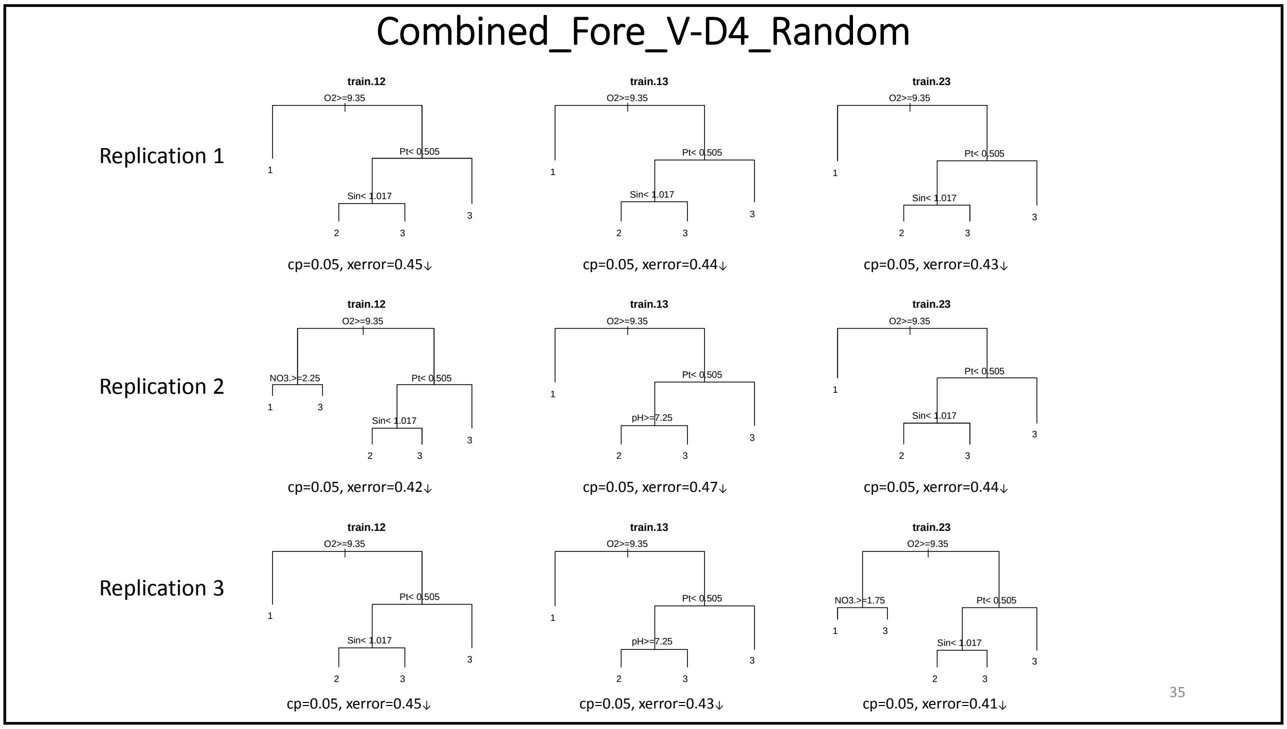
Replication 3

cp=0.05, xerror=0.46↓

cp=0.05, xerror=0.41↓

cp=0.05, xerror=0.44↓

34



Combined_Fore_V-D4_Random

Replication 1

cp=0.05, xerror=0.45↓ cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.43↓

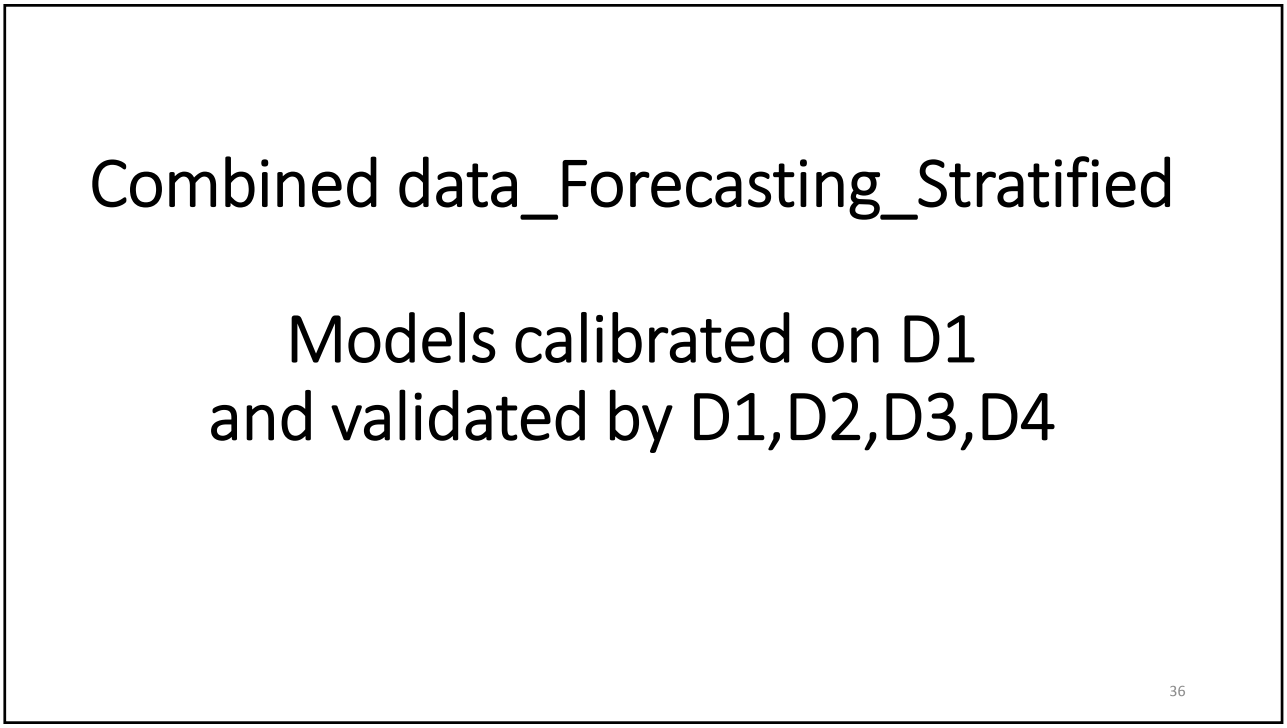
Replication 2

cp=0.05, xerror=0.42↓ cp=0.05, xerror=0.47↓ cp=0.05, xerror=0.44↓

Replication 3

cp=0.05, xerror=0.45↓ cp=0.05, xerror=0.43↓ cp=0.05, xerror=0.41↓

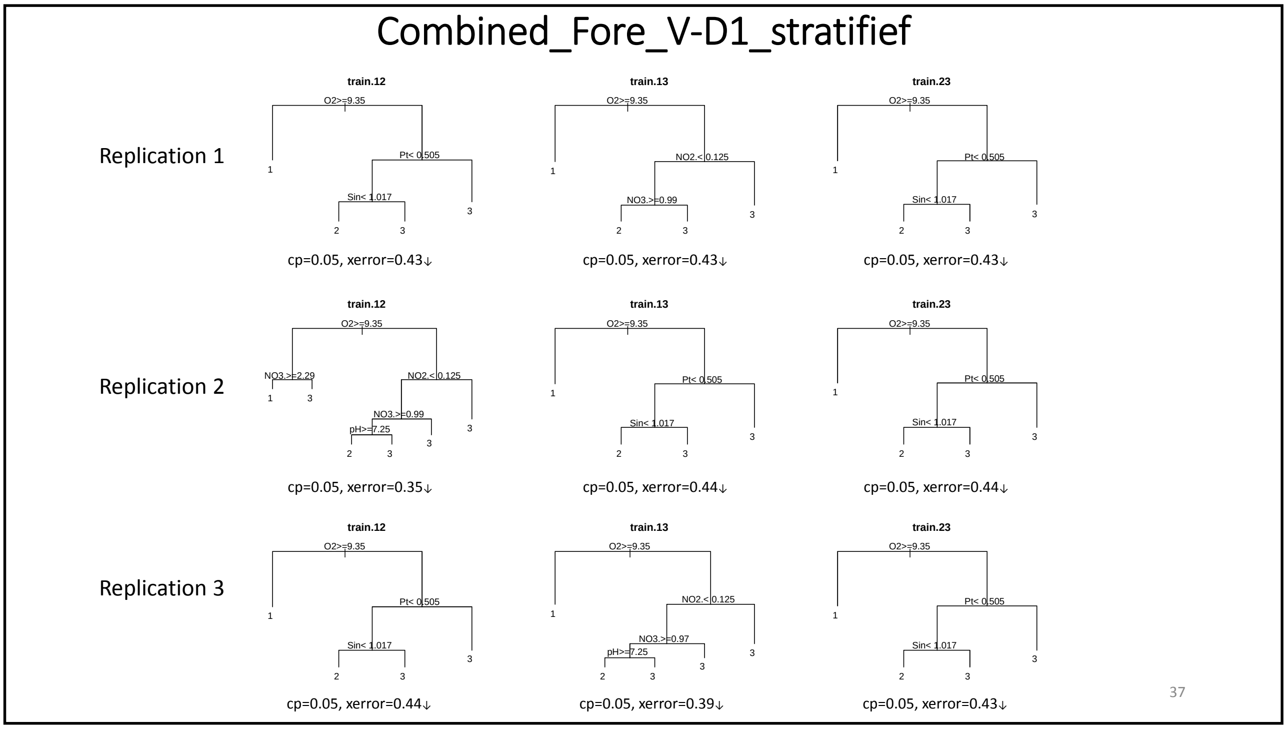
35



Combined data_Forecasting_Stratified

Models calibrated on D1
and validated by D1,D2,D3,D4

36



Combined_Fore_V-D1_stratifief

Replication 1

cp=0.05, xerror=0.43↓ cp=0.05, xerror=0.43↓ cp=0.05, xerror=0.43↓

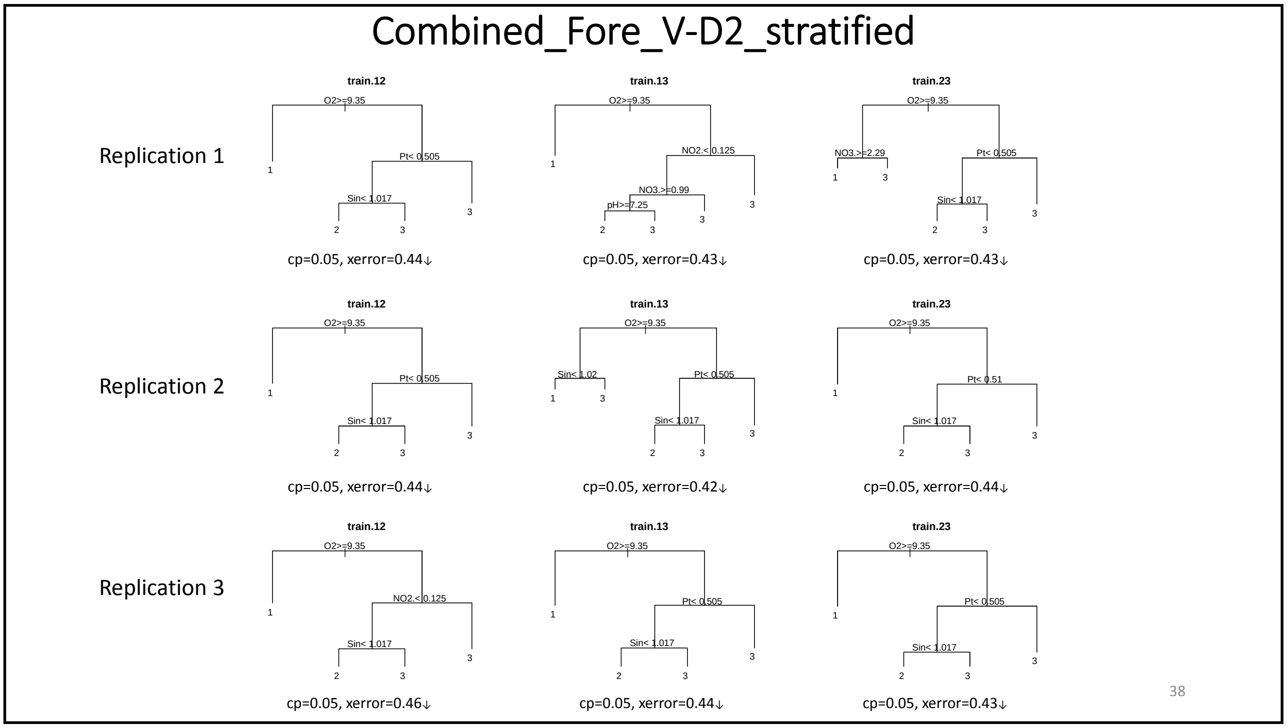
Replication 2

cp=0.05, xerror=0.35↓ cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.44↓

Replication 3

cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.39↓ cp=0.05, xerror=0.43↓

37



Combined_Fore_V-D2_stratified

Replication 1

cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.43↓ cp=0.05, xerror=0.43↓

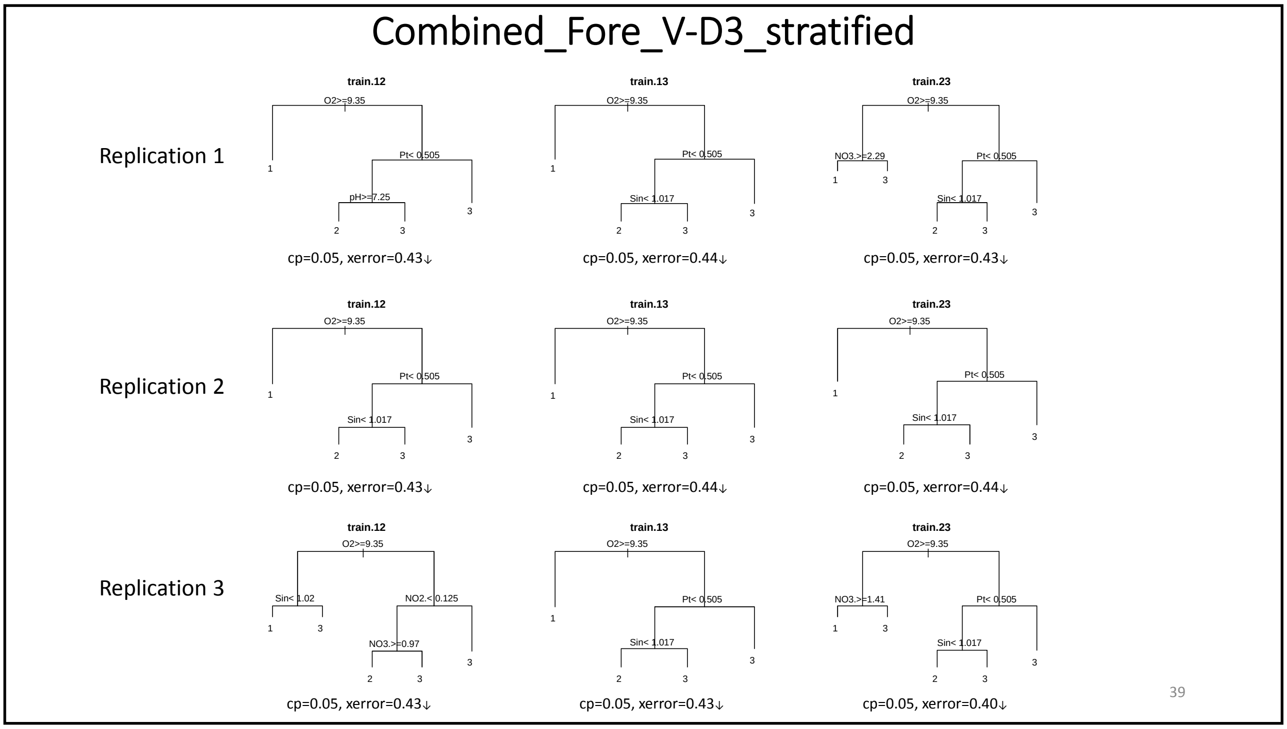
Replication 2

cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.42↓ cp=0.05, xerror=0.44↓

Replication 3

cp=0.05, xerror=0.46↓ cp=0.05, xerror=0.44↓ cp=0.05, xerror=0.43↓

38



Combined_Fore_V-D3_stratified

Replication 1

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.43↓

Replication 2

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.44↓

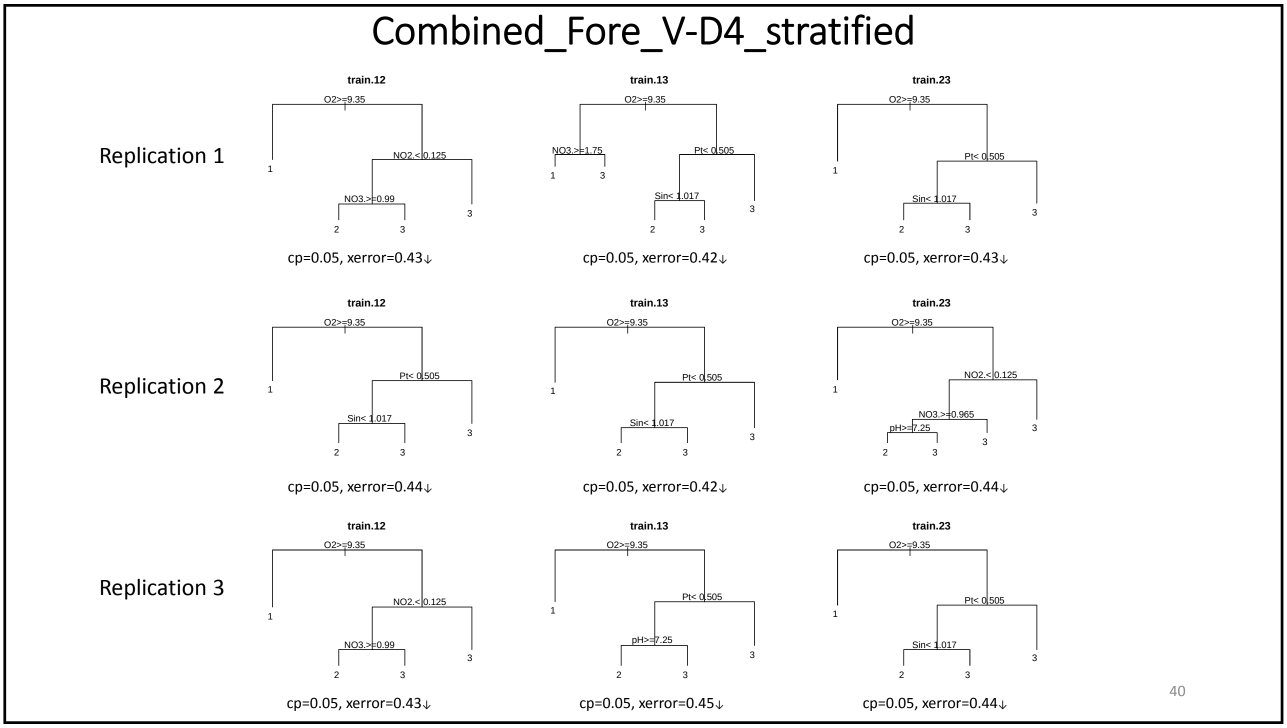
Replication 3

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.40↓

39



Combined_Fore_V-D4_stratified

Replication 1

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.42↓

cp=0.05, xerror=0.43↓

Replication 2

cp=0.05, xerror=0.44↓

cp=0.05, xerror=0.42↓

cp=0.05, xerror=0.44↓

Replication 3

cp=0.05, xerror=0.43↓

cp=0.05, xerror=0.45↓

cp=0.05, xerror=0.44↓

40