Ghent University
Faculty of Bioscience Engineering
Department of Forest and Water Management

Ecohydrology of wetlands: monitoring and modelling interactions between groundwater, soil and vegetation

Ecohydrologie van natte ecosystemen: bemonstering en modellering van interacties tussen grondwater, bodem en vegetatie

Jan Peters

Promotors **Prof. dr. ir. Niko E. C. Verhoest**
Department of Forest and Water Management,
Faculty of Bioscience Engineering, Ghent University

Prof. dr. ir. Roeland Samson
Department of Bioscience Engineering,
Faculty of Science, Antwerp University

Prof. dr. Bernard De Baets
Department of Applied Mathematics, Biometrics and Process Control,
Faculty of Bioscience Engineering, Ghent University

Dean **Prof. dr. ir. Herman Van Langenhove**

Rector **Prof. dr. Paul Van Cauwenberge**

JAN PETERS

ECOHYDROLOGY OF WETLANDS: MONITORING AND MODELLING INTERACTIONS BETWEEN GROUNDWATER, SOIL AND VEGETATION

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (PhD) in Applied Biological Sciences

Dutch translation of the title:
Ecohydrologie van Natte Ecosystemen: Bemonstering en Modellering van Interacties tussen Grondwater, Bodem en Vegetatie

Cover:
Photograph of a flowering Broad-leaved marsh orchid (*Dactylorhiza majalis* subsp. *majalis*), a wonderful but endangered plant of phreatic wetlands.
(Photograph by Staf De Roover)

Reference:
J. Peters. *Ecohydrology of wetlands: monitoring and modelling interactions between groundwater, soil and vegetation*. PhD thesis, Ghent University, September 2008.

ISBN 978-90-5989-000-8

The author and the promotors give the authorization to consult and to copy parts of this work for personal use only. Every other use is subject to the copyright laws. Permission to reproduce any material contained in this work should be obtained from the author.

Dankwoord

De multi-disciplinariteit die een term als ecohydrologie suggereert wordt geïllustreerd door de promotoren die mij steunden tijdens mijn doctoraatsstudie. Prof. Niko Verhoest, hydroloog, Prof. Roeland Samson, ecoloog, en Prof. Bernard De Baets, wiskundige. Ik wens hen alle drie van harte te bedanken voor een fantastische samenwerking. Een selectie uit hun inbreng omvat: het aanleveren van de random forest techniek, het nauwgezet nalezen en (stichtend) becommentariëren van teksten, het voortdurend motiveren... Ik kan niet anders dan besluiten dat dit werk niet tot stand had kunnen komen zonder jullie promotorschap. Waarvoor nogmaals dank!

Vervolgens wens ik Piet De Becker en Willy Huybrechts, beide werkzaam aan het Instituut voor Natuur- en Bosonderzoek, te bedanken voor het aanleveren en toelichten van de ecohydrologische data die gebruikt werden in dit proefschrift. Dave Loete wordt bedankt voor zijn hulp bij het veldwerk en het bemonsteren van het testgebied in de Bourgoyen-Ossemeersen, Bart De Muynck voor zijn hulp bij de vegetatieopnames in de Bourgoyen-Ossemeersen, Rudi Hoeben voor de computer gerelateerde bijstand, en Marcella van Hese voor de logistieke steun. Bedankt Tinne Cockx voor het aanleveren van fotomateriaal. Sven Degroeve, Prof. Pascal Boeckx, Prof. Marc Van Meirvenne, Samuel Verstraete en Liesbet Cockx worden bedankt voor hun wetenschappelijk inbreng. De opmerkingen van de leden van de leescommissie, Dr. Els Ducheyne, Dr. Willy Huybrechts en Prof. Kris Verheyen hebben eveneens bijgedragen tot een verbetering van dit proefschrift, waarvoor dank.

Verder gaat mijn appreciatie uit naar naar de collega's van het Laboratorium voor Hydrologie en Waterbeheer. Wajira, Gabrielle en Lien, en Bruno, Hans en Douglas. Gracias a Jesús Álvarez también. En aan Dave voor straffe verhalen en aangename verpozingen.

Natuurlijk was er ook de fantastische steun van de vrienden, de goalgetters van Sponsor Gezocht (de meest constante ploeg uit het Gentse), een overgelopen Boca Senior, de fietsvrienden van De Lachende Kassei, met onze voorzitter Bert Olivié en secretaris Bert De Jaegher voorop. En het geweldige koppel van around the block, Evert en Veronica, bedankt voor alles... . Ook Klaas, Nacho, Thomas, Juan (Zorro) en Sofieke, Wajira, Udaya en Yoshini, Joke, Johan en Bily, Jan, Katrijn en Klaas, Pim, Anneke, Dirk, Pauwi, Bram, Tinne en Lissa, Klaas en Katrien.

Mijn (schoon)familie, in het bijzonder de zussen Lien, Jet en ons Lotte(ke) en papa Toon en ons moeke, krijgen drie dikke kussen en een dikke merci om altijd en overal zo vrolijk en plezant te zijn!

Maar mijn diepste dankbetuiging gaat naar Liesbet. Zij schonk ons in de loop van mijn doctoraatsstudie een pracht van een Maritje.

Gent, september 2008
Jan Peters

Table of Contents

Dankwoord **i**

Table of Contents **iii**

List of Figures **vii**

List of Tables **xiii**

List of Abbreviations and Acronyms **xv**

List of Notations **xvii**

Samenvatting **xxi**

Summary **xxv**

1 Introducing wetlands and ecohydrology **1-1**
- 1.1 Wetland definition . . . 1-1
  - 1.1.1 Wetland hydrology . . . 1-2
  - 1.1.2 Wetland physicochemical environment . . . 1-4
  - 1.1.3 Wetland biota . . . 1-5
- 1.2 Wetland value and management . . . 1-6
- 1.3 Pattern and scale . . . 1-7
- 1.4 Patterns in the geographical distribution of species . . . 1-9
- 1.5 Ecohydrology: research at the interface between ecology and hydrology . . . 1-9
- 1.6 Summary . . . 1-10

2 Problem definition and research objectives **2-1**
- 2.1 Problem definition . . . 2-1
- 2.2 Research objectives . . . 2-2
- 2.3 Outline . . . 2-3
  - 2.3.1 Research question 1 . . . 2-3
  - 2.3.2 Research question 2 . . . 2-4
  - 2.3.3 Research question 3 . . . 2-4
  - 2.3.4 Research question 4 . . . 2-4

2.3.5 Research questions 5–6 2-4
2.3.6 Research questions 7–8 2-4

3 Description of test sites: monitoring and ecohydrological review 3-1
3.1 Introduction 3-1
3.2 Monitoring of test sites 3-3
3.2.1 Abiotic site characterization 3-3
3.2.2 Biotic site characterization 3-4
3.3 Ecohydrological review of the test sites 3-4
3.3.1 Topography 3-4
3.3.2 Groundwater 3-8
3.3.2.1 Groundwater quantity 3-9
3.3.2.2 Groundwater quality 3-10
3.3.3 Soil solution 3-11
3.3.4 Soil organic matter 3-14
3.4 Summary 3-14

4 Selection and comparison of different vegetation distribution models 4-1
4.1 Introduction 4-1
4.2 Literature overview of distribution models 4-2
4.2.1 Conceptual considerations 4-2
4.2.2 Different distribution models 4-4
4.2.3 Selection of techniques 4-10
4.3 Material and methods 4-11
4.3.1 Study sites 4-11
4.3.2 The ecohydrological data set 4-11
4.3.3 Statistical model description 4-11
4.3.3.1 Multiple logistic regression 4-11
4.3.3.2 Random forests 4-16
4.3.4 Training versus test data sets 4-18
4.4 Model construction, calibration and results 4-19
4.4.1 Multiple logistic regression model 4-19
4.4.2 Random forest model 4-22
4.5 Model evaluation 4-29
4.5.1 Observed versus predicted 4-29
4.5.2 Prediction probabilities 4-34
4.6 Discussion and conclusions 4-38
4.6.1 Statistical model comparison 4-38
4.6.2 Putting the random forest model in a broader perspective . 4-39

5 Identification of important environmental variables in ecohydrological distribution modelling 5-1
5.1 Introduction 5-1
5.2 Material and methods 5-3
5.2.1 Test site and data set 5-3

5.2.2 Detrended correspondence analysis (DCA) and canonical correspondence analysis (CCA) 5-5
5.2.3 Multiple logistic regression 5-5
5.2.3.1 Predictive multiple logistic regression model . . 5-5
5.2.3.2 Hierarchical partitioning 5-6
5.2.4 Random forests 5-6
5.2.5 Data considerations 5-7
5.2.6 Model evaluation statistics 5-11
5.3 Results 5-11
5.3.1 Data inspection 5-11
5.3.2 Ordinations 5-11
5.3.3 Multiple logistic regression model 5-16
5.3.3.1 Predictive multiple logistic regression model . . 5-16
5.3.3.2 Explanatory modelling with hierarchical partitioning 5-16
5.3.4 Random forest model 5-18
5.3.4.1 Construction and evaluation of the best predictive random forest model 5-18
5.3.4.2 Explanatory modelling using the ‘variable importance’ measure 5-21
5.3.5 Predictive random forest modelling on reduced data subsets 5-21
5.3.5.1 Leave-one-variable-out 5-21
5.3.5.2 Gradually decreasing model complexity 5-23
5.3.5.3 Predictive modelling using the selected subset . 5-23
5.4 Statistical model evaluation and discussion 5-24
5.4.1 Predictive modelling 5-24
5.4.2 Explanatory modelling 5-25
5.4.3 Reduced random forest model 5-28
5.5 Conclusions 5-28

6 Independent model testing **6-1**
6.1 Introduction 6-1
6.2 Ecohydrology of the Doode Bemde 6-2
6.2.1 Data set 6-5
6.2.2 Independent evaluation data set 6-5
6.3 Modelling vegetation distributions 6-6
6.3.1 Model construction and results 6-6
6.3.2 Model evaluation 6-6
6.3.2.1 Classification accuracy 6-6
6.3.2.2 Spatially explicit evaluation 6-8
6.3.2.3 Performance on independent test data 6-10
6.4 Conclusions 6-12

7 Assessing uncertainty propagation in the random forest distribution model 7-1

7.1 Introduction . . . 7-1
7.2 Material and methods . . . 7-4
    7.2.1 Study area and data set . . . 7-4
    7.2.2 Variation partitioning in species data . . . 7-5
    7.2.3 Spatial interpolation using sequential Gaussian simulation 7-5
    7.2.4 Species clustering . . . 7-10
    7.2.5 The random forest distribution model . . . 7-10
    7.2.6 Evaluation of distribution modelling results . . . 7-11
7.3 Uncertainty assessment related to uncertainty in environmental variables . . . 7-12
    7.3.1 From field observations to a spatially distributed data set . 7-12
        7.3.1.1 Variation partitioning in species cover data . . . 7-14
        7.3.1.2 Uncertainty on spatial interpolation of environmental variables . . . 7-15
    7.3.2 Model construction, calibration and evaluation . . . 7-19
    7.3.3 Uncertainty on spatial interpolation of environmental variables propagating to the modelling results . . . 7-21
        7.3.3.1 Latin hypercube test data sets . . . 7-21
        7.3.3.2 Deviation from median test data sets . . . 7-24
        7.3.3.3 Implications for empirical distribution modelling 7-26
7.4 Uncertainty assessment related to uncertainty in vegetation distribution . . . 7-26
    7.4.1 Uncertainty on species clustering . . . 7-26
    7.4.2 Uncertainty on species clustering propagating to the modelling results . . . 7-27
7.5 Conclusions . . . 7-29
7.6 Additional remark on the use of time series . . . 7-30

8 Conclusions and recommendations 8-1

8.1 Starting point of this dissertation . . . 8-1
8.2 Answers to the research questions . . . 8-2
8.3 Contribution of this dissertation . . . 8-5
8.4 Future perspectives . . . 8-6
    8.4.1 Compilation of data sets . . . 8-6
    8.4.2 Time dependence . . . 8-7
    8.4.3 Coupled modelling . . . 8-9
8.5 Recommendations for potential users . . . 8-11

Bibliography Bib-1

A Shortlist of plant species A-1

B Photograph of the different vegetation types B-1

List of Figures

1.1 Conceptual model illustrating the three-component basis of a wetland. The three components are not independent, but affecting each other. 1-2
1.2 Wetlands are transitional areas between between upland and aquatic systems. 1-3
1.3 Time sequence of redox transformations after soil submergence. . 1-5
1.4 Conceptual model illustrating the three wetland components (hydrology, physicochemical environment and biota) affected by a fourth component, anthropogenic disturbances. 1-7
1.5 Wetland heterogeneity arises from the interplay between hydrology, physicochemical environment, anthropogenic disturbances and biota acting on a hierarchy of spatial scales. 1-8

3.1 Location of the test sites in Flanders, Belgium. 3-2
3.2 Topography and piezometer (△) locations (a) Doode Bemde, (b) Snoekengracht, (c) Vorsdonkbos-Turfputten and (d) Zwarte Beek. 3-6
3.3 Schematic cross-section of an alluvial floodplain. The gradient in grayscale intensity is illustrative for the textural gradient caused by lateral fining. 3-8
3.4 Water balance of a soil column where the root zone is disconnected from the groundwater table (a), and connected with the water table (b). Explanation of abbreviations: P_n net precipitation, P precipitation, I interception, ET evapotranspiration, S_i and S_o surface inflow and outflow, G_i and G_o groundwater discharge and recharge, $\frac{d\theta(t)}{dt}$ change in volumetric water content (θ) during a period of time t, Z_r depth of root zone, gwt groundwater table. 3-12
3.5 Contour plot of the average groundwater depth [m below soil surface] (a) and soil organic matter content [%] (b) at Doode Bemde. 3-15

4.1 A classification of models based on their intrinsic properties. . . . 4-3
4.2 Conceptualization of distribution models. The pattern in abiotic wetland conditions are used to model species and vegetation distributions assuming a state of equilibrium. 4-5
4.3 Recursive binary partitioning using CART. Partition of a two-dimensional predictor space (a) and the corresponding tree (b). 4-8

4.4 The relationship between a linear combination of the independent variables and the conditional probability that the outcome is present as response variable. 4-12

4.5 Schematic overview of the construction of the random forest model, and its application on a test data set (example of RF1, the construction and testing setup of RF2 is identical). 4-23

4.6 Out-of-bag (oob) error and test set error converge when more trees are added to the random forest. L_{train1} oob error and L_{train2} oob error are the oob errors calculated during the construction of RF1 and RF2 respectively. L_{test1} error and L_{test2} error are the test set error of RF1 and RF2 applied to their respective test data sets. . . 4-25

4.7 Predicted vegetation types with the multiple logistic regression model (a) and with the random forest model (b). The observed vegetation distribution (□) is overlaid with the predicted vegetation distribution (○). For each grid cell, the vegetation type with the highest probability of occurrence, as modelled with the multiple logistic regression model (a) and with the random forest model (b), is the predicted vegetation type. 4-27

4.8 Probability distributions of predictions made with the multiple logistic regression model (a) and the random forest model (b) ($N = 1705$). 4-36

4.9 Example of how a ROC graph is created. Ten test instances of which the presence/absence of a vegetation type is observed and the probability of occurrence is modelled ($P(c_j)$), are used to calculate the true positive rate (tpr) and false positive rate (fpr) of confusion matrices constructed by applying three different threshold probabilities: (A) $P(c_j) = 0$, (B) $P(c_j) = 0.5$, and (C) $P(c_j) = 1$. The ROC curve generated when all possible threshold probabilities are used is presented, and the three (fpr,tpr) pairs calculated are indicated (×). The dashed diagonal line (fpr = tpr) represents the ROC curve when vegetation types are classified by random guessing. The shaded area under ROC curve (AUC) has an area of 0.875, the AUC under the dashed ROC curve equals 0.5. 4-37

5.1 Schematic overview of the construction of the random forest model and the determination of important variables. 5-9

5.2 Graph of the detrended correspondence analysis (DCA) showing the ordination of plant species and vegetation types along the first two ordination axes in relation to the environmental variables. . . 5-14

5.3 Graph of the canonical correspondence analysis showing the position of plant species and vegetation types in relation to the environmental variables. 5-15

5.4 Spatially distributed vegetation types at Doode Bemde. Observations overlaid by predictions made by the logistic regression models (a). Observations overlaid by predictions made by the random forest models (b). 5-18

5.5 Independent individual contributions (I–values) of the variables as determined with hierarchical partitioning. Lower values indicate lower independent importance. 5-19

5.6 Mean importance scores (bars) and *z*–scores (lines) for all 17 variables, for each vegetation type separately and all vegetation types together, as determined with the random forests ‘variable importance’ measure. Lower values indicate lower importance. 5-22

5.7 Out-of-bag (oob) error of models with decreasing number of predictive variables. The order of deletion is taken from results for all vegetation types, as presented in subplot 8 of Fig. 5.6. 5-24

6.1 Vegetation distribution at Doode Bemde with piezometers (○) A–E along a topographical gradient (a) (see also Fig. 3.2). Time series of the groundwater depth, as monitored by piezometers A–E (b). Hydrological duration lines expressing the probability that measured groundwater depths are exceeded. The line colours correspond to the vegetation types wherein these piezometers were installed (c). 6-4

6.2 Observed vegetation types overlaid by the classification made by the random forest distribution model (a). Modelled probabilities ($P(c)_{\text{max}}$) on which the classification is based (b). 6-7

6.3 Receiver operating characteristic (ROC) curves visualizing the classification performances of the 3-fold cross-validated random forest distribution model for the 7 vegetation types (full curves). The $\text{AUC}_{\text{total}}$ equals 0.96. Model performances for boundary cells only are summarized by the dashed ROC curves, yielding an $\text{AUC}_{\text{total}}$ value of 0.92. 6-9

6.4 Probability distribution of correct and incorrect classified grid cells of the Doode Bemde ($N = 519$). 6-10

6.5 Conceptual representation of realised niches of *Calthion palustris* at the Doode Beemde and Snoekengracht. The fundamental niche of *Calthion palustris* ranges over all environmental states which would permit to *Calthion palustris* to exist indefinitely. 6-12

7.1 Assessment strategy for uncertainty propagation due to uncertainty associated with the spatial interpolation of environmental variables (a), and associated with species clustering into vegetation types (b). The latter was investigated by strategy (b2), (b1) was not applied. 7-3

7.2 The construction of spatial variables starts from spatial information (X and Y coordinates) which is used to calculate Euclidean distances. Principle coordinate analysis of the truncated distance matrix (with a given maximal distance (max)) results in a number of positive eigenvaulues which are used as spatial variables in a direct gradient analysis. 7-6

7.3 Variation partitioning showing the 4 different fractions. 7-7

7.4 Flowchart of sequential Gaussian simulations for groundwater depth at Doode Bemde. 7-8

7.5 Groundwater depths were monitored by piezometers (black dots, $n = 24$) scattered in and around the study area (a). Sequential Gaussian simulation using these observations resulted in 500 equiprobable groundwater depth realizations for each grid cell ($N = 519$). Empirical non-parametric conditional cumulative distribution functions (ccdfs) were computed from these realizations (b). Median (c) and variance (d) values were calculated based on the unique ccdf of each grid cell. 7-17

7.6 Scatter diagram (1), accuracy plot (2) and precision plot (3) for the simulation results of the jackknifed elements of AGD (a), Ampli (b), pH (c), Cl^- (d), SO_4^{2-} (e), SOM (f). 7-18

7.7 Out-of-bag (oob) error and test set error converge when more trees are added to the random forest (when k increases). The numbers of variables (m) used to split the nodes are $m = 1$, $m = 3$ and $m = 7$. Average error values of the 3-fold cross-validated random forest model are plotted. 7-20

7.8 Histogram of frequency counts of the Shannon entropy (H) values of the entire study site ($N = 519$) for the random forest distribution model cross-validated on L (a), and tested on the Latin hypercube samples (averaged) (b), and the deviation from median test data sets $L_{m+a\times stdev}$ with $a = 0.01$, $a = 0.1$, and $a = 1$ (c). 7-22

7.9 Variable variance versus variance in modelled probability of occurrence of the predicted vegetation type ($P(c)_{\max}$) when the random forest distribution model is applied to 100 Latin hypercube test data sets. Different colours group the scatter points ($N = 519$) based on classification accuracy: a turquoise × is a grid cell classified correctly in ≤ 25 Latin hypercube model testing runs (out of 100), a green × is a grid cell classified correctly in > 25 and ≤ 50 model testing runs, a blue × is a grid cell classified correctly in > 50 and ≤ 75 model testing runs, and a black × is a grid cell classified correctly in > 75 and ≤ 100 model testing runs. Spearman's rank correlations (r_s) and significance at the 0.05 significance level (*) are indicated for each group separately. 7-25

7.10 Cluster dendrogram. 7-27

7.11 The uncertainty of the individual groundwater depth measurements ($u(z_i)$) and uncertainty in function of incomplete time coverage for sites with different groundwater depth dynamics (represented by standard deviations $\sigma(\hat{\mathbf{z}})$). 7-32

8.1 Conceptual representation of the time dependence of vegetation responses to environmental changes. Two vegetation types have unimodal responses to an environmental gradient, showing an optimum, however, at different environmental states. An environmental state shift from A to B results in a decrease of the modelled probability of occurrence for vegetation type 1 ($P_{A,1} \rightarrow P_{B,1}$), while it increases for vegetation type 2 ($P_{A,2} \rightarrow P_{B,2}$). Predictions of a stationary model (crosses in right panel) do not give any information about the time-scale on which vegetation changes are occurring in response to change in environmental change. Different trajectories are possible (tr_1, tr_2 and tr_3, indicated by the green dashed line). 8-8

8.2 Framework for coupled modelling. 8-10

B.1 The forest type *Alno - Padion* is present at Snoekengracht (April 2008). B-1

B.2 *Arrhenatherion elatioris* with several grass species, as observed at Bourgoyen-Ossemeersen, Gent. B-2

B.3 *Caltha palustris* is a diagnostic species for the *Calthion palustris* vegetation type. The blueish aspect of the water on the soil surface is characteristic for seepage water. B-3

B.4 *Carici elongatae – Alnetum glutinosae* at Snoekengracht (April 2008). An *Alnus glutinosa* tree layer is undergrown by a herblayer with *Carex acutiformis*, as can be seen on the foreground of this photograph. B-4

B.5 *Caricion curto-nigrae* with flowering *Lychnis flos-cuculi* at Zwarte Beek has a high species similarity with *Calthion palustris*. B-5

B.6 Two adjacent vegetation types at Vorsdonkbos-Turfputten: the herbaceous *Cirsio – Molinietum* on the foreground and *Sphagno – Alnetum*, the woody vegetation type on the background. B-6

B.7 *Filipendula ulmaria* is a diagnostic species for the *Filipendulion* vegetation type. B-7

B.8 *Magnocaricion* vegetation type, with detail of a flowering *Carex* species. B-8

B.9 *Magnocaricion* with *Phragmites* at Snoekengracht (foreground, April 2008). B-9

B.10 The brown *Phragmitetalia* vegetation belt (winter aspect) at Snoekengracht (April 2008). B-10

List of Tables

3.1 Overview of the test sites: name, abbreviation (Abbr.), location, area, grid size and number of grid cells (Nr.). 3-3

3.2 Summary of the vegetation types: number, name, short description and area. 3-5

4.1 Exampes of link functions for several distributions. 4-6

4.2 Overview of generalized linear models (GLM) and generalized additive models (GAM) in distribution modelling with information on the response variable (type, probability distribution and examples), the model prediction and literature examples. No case studies found is indicated by —. 4-7

4.3 Overview of distribution models using tree-based techniques, canonical correspondence analysis (CCA), redundancy analysis (RDA), Bayesian techniques, artificial neural networks (ANN), support vector machines (SVM), and random forest (RF). 4-8

4.4 Translation of a categorical predictive variable into design variables using Helmert contrasts. 4-13

4.5 Instances with equal measurement vectors are grouped to determine Pearson residuals and Pearson goodness-of-fit. 4-15

4.6 Model goodness-of-fit. 4-21

4.7 Oob error values for RF1 and RF2 built on L_{train1} and L_{train2} respectively. Test set error values for RF1 and RF2 applied to L_{test1} and L_{test2} respectively. 4-25

4.8 Model performances represented by confusion matrices in which observed vegetation types are compared with predicted vegetation types using the multiple logistic regression model (a) and the random forest model (b). 4-30

4.9 Contingency table for the McNemar test. 4-31

4.10 Confusion matrix. TP stands for True Positive, FP for False Positive, FN for False Negative and TN for True negative. 4-31

4.11 McNemar test for comparison of MLR and RF model performances for each vegetation type individually. Precision, recall, and three F-measures ($F_{0.5}$, F_1, F_2) for MLR and RF modelling results are given for each vegetation type separately as well. . . . 4-33

4.12 Area under ROC curves for the MLR and the RF model. 4-38

5.1 Summary of the quantitative environmental variables as measured at the Doode Bemde. All values in [mg L^{-1}] except for average groundwater depth (AGD) [m], maximal groundwater depth (Max) [m], minimal groundwater depth (Min) [m], amplitude (Ampli) [m], pH [-] and soil organic matter content (SOM) [%]. The categorical variables soil type (loam, peat) and management (yearly, transition from yearly to cyclic (Y/C), cyclic, no management (no man.)) have no mean and no standard deviation [/]. Kendall correlation coefficients (τ) of the first two DCA and CCA axes are included. Four environmental variables were excluded from the CCA due to multicollinearity problems [–]. 5-4
5.2 Kendall τ correlations between quantitative variables. 5-12
5.3 Summary statistics for the DCA and CCA ordinations. 5-13
5.4 Goodness-of-fit of the multiple logistic models. Residual deviances ($D_{\text{resid.}}$) and Akaike's information criterion (AIC) for the different vegetation types. 5-17
5.5 Out-of-bag (oob) error values and test set error values for RF_{12}, RF_{13} and RF_{23} constructed with several values of m (m is the number of randomly sampled variables to split the nodes). 5-20
5.6 Kendall τ correlations between results of hierarchical partitioning (I–values) and 'variable importance' (mean importance scores). . 5-25

6.1 Jaccard index of similarity between the vegetation types in the Doode Bemde. Vegetation types are abbreviated accordingly the List of Abbreviations and Acronyms. 6-3
6.2 Confusion matrix of the classification made by the random forest distribution model. Predicted vegetation types are compared with the observations at the Doode Bemde. 6-6
6.3 Jaccard index of similarity for boundary grid cells between two vegetation types at the Doode Bemde. Non-adjacent vegetation types are indicated by a dash. 6-11

7.1 Spatio-temporal resolution of field observations made within different ecosystem compartments. Derived variables, abbreviations and summary statistics are included. 7-13
7.2 Summary of semivariogram models. 7-15
7.3 Uncertainty on environmental variables propagating to the random forest distribution modelling results. 7-24
7.4 Uncertainty on species clustering propagating to the random forest distribution modelling results. 7-28

List of Abbreviations and Acronyms

DB	Doode Bemde
SN	Snoekengracht
VB	Vorsdonkbos-Turfputten
ZB	Zwarte Beek

AP	*Alno – Padion*
Ar	*Arrhenaterion elatioris*
Cp	*Calthion palustris*
Cc	*Caricion curto – nigrae*
Ce	*Carici elongetae – Alnetum glutinosae*
CM	*Cirsio – Molinietum*
Fi	*Filipendulion*
Ma	*Magnocaricion*
MP	*Magnocaricion* with *Phragmites*
Ph	*Phragmitetalia*
SA	*Sphagno – Alnetum*

AGD	average groundwater depth
Ampli	amplitude of the groundwater depth
IR	ionic ratio
Max	maximal groundwater depth
Min	minimal groundwater depth
SOM	soil organic matter

ANN	artificial neural network
CCA	canonical correspondence analysis
ccdf	conditional cumulative distribution function
DCA	detrended correspondence analysis
GAM	generalized additive model
GLM	generalized linear model
MLR	multiple logistic regression
oob	out-of-bag
PCA	principle component analysis

PCoA	principle coordinate analysis
RDA	redundancy analysis
RF	random forest
sGs	sequential Gaussian simulation
SVM	support vector machine
TWINSPAN	two way indicator species analysis

List of Notations

x	scalar
$\mathbf{x}$	vector
X	matrix

L	data set
L_{train}	training data set
L_{test}	test data set
L_{ev}	independent evaluation data set
C	collection of vegetation types
c_j	vegetation type j, $c_j \in C$
$\mathbf{x}_i$	measurement vector of grid cell i
l_i	vegetation type of grid cell i, $l_i \in C$
N	number of elements
n	number of observations

$\pi(\mathbf{x})$	conditional probability given $\mathbf{x}$
$\hat{\pi}(\mathbf{x})$	estimated conditional probability given $\mathbf{x}$
$g(\mathbf{x})$	link function
$\hat{g}(\mathbf{x})$	estimated link function
$l(\beta)$	likelihood function
$L(\beta)$	log-likelihood function
k	number of trees in the random forest
m	number of predictive variables to split nodes
$P(c_j)$	probability of occurrence of vegetation type c_j
$P(c)_{\text{max}}$	maximal probability of occurrence
$E(\cdot)$	expected response

Z	random variable
z	attribute value of random variable Z
$\mathbf{u},\mathbf{v}$	location vectors
l	conditioning information
p	probability, $p \in [0,1]$
$P(Z)$	probability distribution of Z

$P(Z\|l)$	conditional probability of Z given l
$F(\cdot\|l)$	conditional cumulative distribution function
$F^{-1}(p)$	inverse conditional cumulative distribution function
$F_{\mathbf{u}}(\cdot\|l)$	conditional cumulative distribution function at location $\mathbf{u}$
$F_{\mathbf{u}}^{-1}(p)$	inverse conditional cumulative distribution function at location $\mathbf{u}$
$\hat{F}(\cdot\|l)$	estimated conditional cumulative distribution function
$\hat{F}^{-1}(p)$	estimated inverse conditional cumulative distribution function
$G(\cdot)$	Gaussian cumulative distribution function
$G^{-1}(\cdot)$	inverse Gaussian cumulative distribution function
$\xi(\cdot)$	step function
$w(\cdot)$	weighing factor
h	lag distance
$\gamma(h)$	semivariogram model
$\hat{\gamma}(h)$	estimated semivariogram model
μ	mean
m	median
$stdev$	standard deviation
$u(z_i)$	uncertainty on measurement z_i
$u(\bar{\bar{\mathbf{z}}})$	uncertainty due to averaging
$u_{\mathrm{TC}}(\bar{\bar{\mathbf{z}}})$	uncertainty on average due to incomplete time coverage

AIC	Akaike information criterion
AUC	area under curve
D	deviance
Df	degree of freedom
F	Van Rijsbergen F-measure
FN	false negative
FP	false positive
fpr	false positive rate
G	likelihood ratio
H	Shannon enthropy
JS	Jaccard similarity index
κ	Cohen kappa
MAE	mean absolute error
M	McNemar test statistic
p	precision
r	recall
r	Pearson (linear) correlation coefficient
r^2	coefficient of determination

r_s	Spearman rank correlation coefficient
RMSE	root mean square error
ROC	receiver operating characteristic
τ	Kendall correlation coefficient
TN	true negative
TP	true positive
tpr	true positive rate
X_p^2	Pearson chi-square goodness-of-fit

Samenvatting

Waterrijke gebieden (*wetlands*) zijn periodiek of permanent nat door hun ligging in het landschap. Deze periodieke of permanente natte status initieert tal van chemische, fysische en biologische processen, die karakteristiek zijn voor deze ecosystemen, en waardevol voor de samenleving. Helaas worden wetlands vaak onbehoorlijk beheerd, waardoor ze degraderen en verloren gaan. Bijgevolg worden ze bij de meest bedreigde ecosystemen gerekend.

Dit proefschrift omvat een ecohydrologische studie van wetlands, met bijzondere aandacht voor het bemonsteren en modelleren van de interactie tussen grondwater, bodem en vegetatie. Het uiteindelijke doel van deze studie is om een instrument te voorzien, dat kan ingezet worden bij het beheer van natte ecosystemen, doordat het de mogelijkheid biedt om de vegetatierespons op veranderende milieucondities te voorspellen, of omgekeerd, doordat het de mogelijkheid biedt om gerichte beheerstaken te formuleren om wetlandvegetaties te herstellen en te conserveren.

Vier waterrijke, alluviale gebieden werden door het Instituut voor Natuur- en Bosonderzoek bemonsterd met betrekking tot verschillende milieuvariabelen gerelateerd aan de grondwaterkwantiteit en -kwaliteit, bodem, en vegetatiebeheer. Een ruimtelijke interpolatie resulteerde in een gridsgewijze, gebiedsdekkende schatting van deze milieuvariabelen. Aanvullend werd ook het voorkomen van plantensoorten op dezelfde gridsgewijze, gebiedsdekkende manier geïnventariseerd. Op basis hiervan werden vegetatietypes gedefinieerd en afgebakend binnen de studiegebieden, wat leidde tot een gebiedsdekkende vegetatiekaart. De sterke interacties tussen het abiotisch milieu enerzijds, en de vegetatiedistributie anderzijds, bieden de mogelijkheid om de distributie van wetlandvegetaties te voorspellen op basis van de abiotische milieucondities in voorspellende vegetatiedistributiemodellen.

In dit proefschrift werd een *ensemble learning* techniek, *random forest*, geïmplementeerd in een voorspellend vegetatiedistributiemodel voor natte ecosystemen. Een logische opeenvolging van onderzoeksvraagstukken met betrekking tot die implementatie werd beantwoord doorheen deze studie.

Vooreerst werd onderzocht of de random forest techniek kon gebruikt worden in een vegetatiedistributiemodel. Daartoe diende voldaan te worden aan een aantal vereisten met betrekking tot het gebruik van continue en categorische milieuvariabelen, de interpretatie van modelresultaten als een waarschijnlijkheid van voorkomen van de verschillende vegetatietypes, een objectieve vergelijking van modelresultaten om te komen tot een finale vegetatievoorspelling, en het invoeren van modelresultaten in een geografisch informatiesysteem. Aangezien de random forest techniek voldeed aan al deze voorwaarden, werd de techniek geïmplementeerd in een distributiemodel, het random forest distributiemodel, dat in een volgende fase van het onderzoek geëvalueerd werd op zijn voorspellende capaciteit.

Daartoe werd een andere techniek, logistische regressie, geselecteerd ter vergelijking. Beide technieken werden geïmplementeerd in twee verschillende distributiemodellen, respectievelijk het logistische regressiemodel en het random forest model. Beide modellen werden gecalibreerd en gekruisvalideerd. De voorspelde vegetatiedistributies werden vergeleken met de geobserveerde vegetatiedistributies. Het logistisch regressiemodel maakte voor 69.3% van de gridcellen een correcte voorspelling, terwijl het random forest model voor 76.7% correct scoorde. De McNemar test gaf een significant betere modelvoorspelling door het random forest model ($p < 0.001$). Wanneer de modelvoorspellingen vergeleken werden voor elk vegetatietype afzonderlijk, met behulp van de F-maat, werd de significant betere prestatie van het random forest model bevestigd ($p = 0.003$).

Wanneer de kans van voorkomen van elk vegetatietype per gridcel vergeleken werd, werd geconstateerd dat correcte voorspellingen in het centrale gedeelte van homogene vegetatieclusters vaak gebaseerd waren op een hoge voorspelde kans van voorkomen, terwijl deze afnamen naar de grenzen van deze gebieden. Het algemene besluit van dit onderzoeksdeel was dat het gebruik van de random forest techniek tot betere distributiemodellen kan leiden.

Een goede wetenschappelijke kennis van wetlands is onontbeerlijk voor het definiëren van correcte en gerichte beheersmaatregelen. Hierbij is kennis omtrent vegetatiedistributies in relatie tot milieugradiënten erg belangrijk. Niet alle milieugradiënten zijn echter even determinerend met betrekking tot vegetatiedistributies. Sommige hebben enkel een indirect effect, andere een direct fysiologisch effect. De milieugradiënten opgenomen in dit proefschrift waren grondwaterkwantiteit en -kwaliteit, bodem en beheer gerelateerd, en er werd verondersteld dat deze gradiënten niet allemaal dezelfde invloed hadden op de vegetatiedistributie. Daartoe werd een onderzoeksdoelstelling geformuleerd om de belangrijkste milieugradiënten te identificeren, gebruik makende van recent ontwikkelde methodes, zoals een hiërarchische partitie van de modelresultaten van het logistisch regressiemodel en een maat voor de belangrijkheid van variabelen die in het random forest algoritme is opgenomen. Uit een vergelijkende studie bleek dat de verschillende methodes verschillende gradiënten identificeerden als zijnde belangrijk, wat hun toepasbaarheid in vraag stelt. Niettegenstaande deze tekortkoming, werden random forest modellen geconstueerd met afnemende complexiteit. Daarbij werden modelvariabelen stapsgewijs geschrapt, te beginnen bij de minst belangrijke, zoals berekend door het random forest algoritme. De modelprestaties waren niet significant verschillend op het 0.05 significantieniveau ($p = 0.016$) wanneer het model gebaseerd was op alle 17 of op slechts de 6 meest belangrijke milieuvariabelen, en alle modelcomplexiteiten daartussen. Daarom werd er geconcludeerd dat het random forest distributie model afdoende presteerde, zelfs bij een sterk gereduceerd aantal milieuvariabelen.

In een volgend deel werd een vegetatiedistributiemodel geconstrueerd voor één van de studiegebieden, gebruik makende van alle milieuvariabelen. Voor elke gridcel werd de kans van voorkomen gemodelleerd, en beoordeeld op het voorkomen van ruimtelijke trends. De gemodelleerde kans van voorkomen was significant lager voor gridcellen met minstens één ander aangrenzend vegetatietype, wat resulteerde in slechtere modelprestaties voor deze gridcellen.

Hetzelfde model werd vervolgens toegepast op een zeer gelijkaardig wetland, om de algemene toepasbaarheid van het model op een onafhankelijk gebied te kunnen inschatten. Met een correcte modelvoorspelling voor slechts 19.8% van de gridcellen, kon besloten worden dat het random forest model niet toepasbaar was buiten het gebied waarop het geconstrueerd werd, omdat de gerealizeerde niche van vegetatietypes zelden volledig overlapt, zelfs in twee zeer gelijkaardige gebieden. Om het model toepasbaarbaar te maken op grotere schaal, dient de volledige ecologische amplitude van de verschillende vegetatietypes opgenomen te worden.

Het laatste deel van dit proefschrift behandelt de onzekerheden gerelateerd aan vegetatiedistributiemodellering. Onzekerheden komen voort uit gegevensbeperkingen, die kunnen veroorzaakt worden door meetfouten, systematische afwijkingen in de meetapparatuur, het verwaarlozen van belangrijke milieugradiënten, of een ruimtelijk of temporeel tekort aan observaties om de lokale variabiliteit te kunnen inschatten. Daarenboven introduceert het model zelf onzekerheid, door het onvermogen om de complexiteit van de ecologische processen die aan de basis liggen van de vegetatiedistributie volledig te vatten. Tenslotte is de modelevaluatie niet vrij van onzekerheid. Twee belangrijke bronnen van onzekerheid werden hieruit gelicht, namelijk de onzekerheid geassocieerd met ruimtelijke interpolatie van milieuvariabelen en de onzekerheid geassocieerd met het groeperen van plantensoorten in vegetatietypes.

Een entropie-gebaseerde onzekerheidanalyse werd uitgevoerd, en de lokale onzekerheid geassocieerd met ruimtelijke interpolatie van puntgegevens werd begroot gebruik makend van sequentiële Gaussiaanse simulaties. De modelresultaten gaven duidelijk aan dat deze bron van onzekerheid zich voortzet naar de modelresultaten. Bemonsteringsprotocols voor natte ecosystemen zouden geconditioneerd kunnen worden om deze bron van onzekerheid te verkleinen. Pseudorandomisaties werden uitgevoerd om het effect van een onzekere afbakening van vegetatietypes te begroten. Modelresultaten gaven duidelijk aan dat ook deze bron van onzekerheid zich verderzet, waardoor het belang van een correcte soortengroepering benadrukt werd.

Tenslotte werden, op basis van de onderzoeksresultaten van dit proefschrift, onderzoeksperspectieven geformuleerd die kunnen leiden tot een verdere verbetering van distributiemodellen.

Summary

Wetlands are land areas that are periodically or permanently wet due to their location in the landscape. The periodical or permanent presence of wet conditions trigger chemical, physical and biological processes that are unique for wetlands. These characteristic processes resulted in the recognition of wetlands as multi-functional areas providing many commodities and values to human society. Unfortunately, wetland management is frequently inappropriate, leading to wetland degradation and loss. Consequently wetlands are ranked among the most threatened ecosystems worldwide.

This dissertation comprises an ecohydrological wetland study, with emphasis on monitoring and modelling of the interaction between groundwater, soil and vegetation. The ultimate goal is to provide a tool which can be implemented for wetland management by enabling the prediction of vegetation responses on environmental changes, and inversely, by enabling the determination of appropriate wetland management tasks to restore and conserve wetland vegetation.

Four alluvial wetlands were monitored by the Research Institute for Nature and Forest on several abiotical environmental variables related with groundwater quantity and quality, soil, and vegetation management. Spatial interpolation resulted in area covering grid estimates of these environmental gradients within the study sites. Additionally, plant species occurrences were mapped using the same grid, and clustered into discrete vegetation types in area covering vegetation maps. The strong linkages between the abiotical wetland environment and wetland vegetation distributions, produce the ability to predict the wetland vegetation distribution based on the distribution of environmental wetland variables in predictive vegetation distribution modelling.

In this dissertation a recently developed ensemble learning technique called ‘random forest’ was implemented for wetland vegetation distribution modelling based on hydrological, hydrochemical, soil and anthropogenic wetland features. A sequence of research questions associated with the implementation of the random forest distribution model was addressed throughout the dissertation.

Firstly, it was investigated whether the random forest technique could possibly be used within a distribution modelling context. Therefore, several requirements should be satisfied, such as the ability to cope with continuous and categorical environmental variables to model the vegetation distributions upon, the ability to interpret the model output as a probability of occurrence for several vegetation types, the ability to compare the model output over different vegetation types to get an objective final prediction, and the ability to incorporate the final prediction into a geographical information system. Since the random forest technique satisfied these requirements, it was implemented in a distribution model, described as the random forest distribution model, in a subsequent research phase focusing on the model's predictive ability.

Another technique, multiple logistic regression, was selected for comparative reasons. Both techniques were included in two separate wetland vegetation distribution models, the multiple logistic regression model and the random forest model, respectively. After model construction and calibration, both models were applied to an independent ecohydrological test data set, including spatially distributed information on several environmental variables related with wetland hydrology, hydrochemistry, soil, and management, using cross-validation. Vegetation distributions, as predicted by both models, were compared with site observations. The multiple logistic regression model made correct predictions in 69.3% of all cases, whereas the random forest distribution model in 76.7% of all cases. A McNemar test indicated a significant better performance of the random forest distribution model ($p < 0.001$). Comparison of the modelling results for each vegetation type seperately by means of the F-measure also revealed a significant better performance of the random forest distribution model ($p = 0.003$).

Inspection of the probabilities of occurrence of the different vegetation types for each grid cell demonstrated that correct predictions in central areas of homogeneous vegetation sites were based on high probabilities, whereas the confidence decreased towards the margins of these areas. The overall conclusion of the predictive ability assessment was that the inclusion of the random forest technique has the ability to lead to better distribution model performances.

Wetland ecosystems are of primary concern for nature conservation and restoration. Adequate conservation and restoration strategies emerge from a scientific comprehension of wetland properties and processes. Hereby, the understanding of vegetation distributions in relation to environmental gradients is an important issue. The multiple logistic regression and random forest modelling approaches relate wetland vegetation distribution to measured environmental gradients statistically. However, not all environmental gradients have the same degree of causality on vegetation distributions, some have an indirect impact whereas others have a direct physiological impact. The environmental gradients included in this dissertation were groundwater quantity and quality aspects, soil properties and vegetation management related, and it was hypothesized that not all gradients were constraining vegetation distributions equally. Therefore, a research objective was formulated to identify the key environmental gradients constraining the vegetation, using recently developed methodologies, hierarchical partitioning of the goodness-of-fit of multiple logistic regression models with gradually increasing complexity and the variable importance measure within the random forest model. Comparison of results indicated that different environmental gradients were considered to be important in constraining vegetation distributions by different methodologies, limiting the applicability of these methodologies. Notwithstanding this drawback, a performance assessment of random forest distribution models with reduced complexity was made based on the variable importance ranking. Model performances were not significantly different ($p = 0.016$) at the 0.05 significance level for model complexities ranging between the full model, based on all 17 environmental variables, and the reduced model using only the 6 most important environmental variables. This assessment allowed to conclude that, despite a methodology dependent variable importance ranking, the prediction of vegetation types based on environmental gradients was satisfactory even if a reduced number of gradients were included.

In the next part of the study, a wetland vegetation distribution model was constructed for one of the test sites using all environmental variables. For each grid cell included, a probability of occurrence value was modelled and assessed on spatial trends. Significantly lower probability values were prevalent for boundary grid cells, i.e. grid cells for which at least one of the adjacent grid cells has a distinct vegetation type, and resulted in higher prediction errors for these areas.

The same model was then applied to an ecologically similar but distant wetland, to assess the generalization ability of the model to fully independent sites. From the 501 grid cells included in the independent test data set, only 99 elements were classified correctly (19.8%). The random forest distribution model could not be applied beyond the local conditions upon which it was constructed, because realized niches of vegetation types do seldom coincide, even between apparently similar sites, hence restricting the model's applicability. It was concluded that in order to make the model operational on a larger scale many data would be needed, ranging over the entire ecological amplitude of the modelled vegetation types.

The last part of this dissertation covers uncertainty aspects associated with wetland vegetation distribution modelling. Uncertainty originates from input data limitations, caused by measurement errors on observations, bias in measurement equipment, neglecting key environmental variables, or a spatial and temporal underrepresentation of observations to capture local variability. Furthermore, the model itself introduces uncertainty due to its disability to capture the entire complexity of ecological processes in relation to vegetation distributions. Finally, model evaluation is also susceptible to uncertainty. Among this variety of uncertainty, focus was exclusively on two sources, namely, the uncertainty associated with the spatial interpolation of environmental variables where predicted vegetation distributions are based upon and the uncertainty associated with species clustering into vegetation types.

An entropy-based uncertainty assessment was set up, and the local uncertainty associated with spatial interpolation of environmental point measurements was quantified using sequential Gaussian simulation. This source of uncertainty clearly propagated toward the random forest distribution modelling results, conditioning monitoring protocols to lower this source of uncertainty. Pseudo-randomizations were performed to quantify the uncertainty propagation associated to species clustering. A deterioration of the modelling results stressed the importance of accurate species clustering.

Finally, based on the results of the sequence of research steps addressed in this dissertation, future perspectives aiming for the improvement of distribution models were formulated.

1

Introducing wetlands and ecohydrology

Among the enormous variety of natural and human ecosystems, this study concentrates on wetlands. Wetlands are land areas that are periodically or permanently wet due to their location in the landscape. They are frequently transitional between upland and aquatic ecosystems. The (periodically) wet conditions trigger chemical, physical and biological processes that are unique to wetland ecosystems. Many wetland commodities and wetland values arise from these wetland processes, and they should therefore be covered by wetland sciences and wetland ecohydrological research.

Given the fact that wetlands exhibit characteristic properties and processes which are fundamental to the objectives and outline of this dissertation, a brief introduction to wetlands and ecohydrology is given previous to the problem definition and research objectives (Chapter 2) of this study. Emphasis is on the construction of a conceptual framework in which wetlands are generally defined by their predominant components, facilitating the interpretation of the results presented throughout this dissertation.

1.1 Wetland definition

Wetland definitions often include three main components [1]:

(1) Wetlands are distinguished by the presence of water, either at the surface or

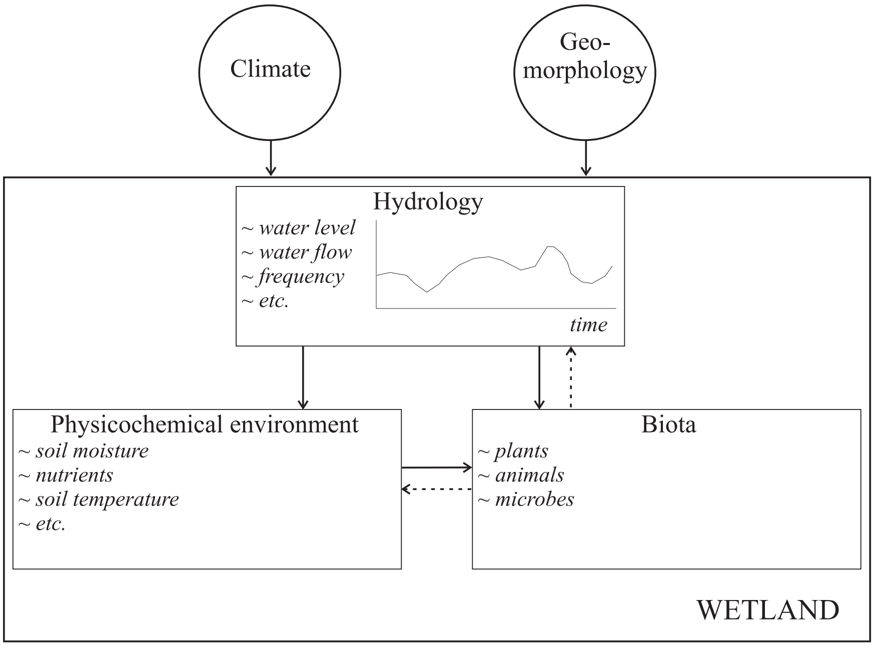


FIGURE 1.1 – Conceptual model illustrating the three-component basis of a wetland (modified from [1]). The three components are not independent, but affecting each other. Legend: Full arrows indicate direct effect, broken arrows feedback from biota.

within the root zone.

(2) Wetlands often have unique soil conditions that differ from adjacent upland and aquatic systems.

(3) Wetlands support biota adapted to wet conditions and, conversely, are characterized by an absence of flooding-intolerant biota.

This three-level definition is reflected in Fig 1.1. The environmental determinants climate and geomorphology define the degree to which wetlands can exist, but the starting point is *hydrology*, which, in turn, affects the *physicochemical environment*, which, in turn, determines together with the hydrology what and how much *biota* inhabit the wetland. Biota, in turn, affect the hydrology and physicochemical environment of the wetland.

1.1.1 Wetland hydrology

Wetlands are transitional areas between upland and aquatic ecosystems (Fig. 1.2) and form an aquatic boundary to many terrestrial plants and animals, and they also form the terrestrial boundary of many aquatic plants and animals [1]. They are also

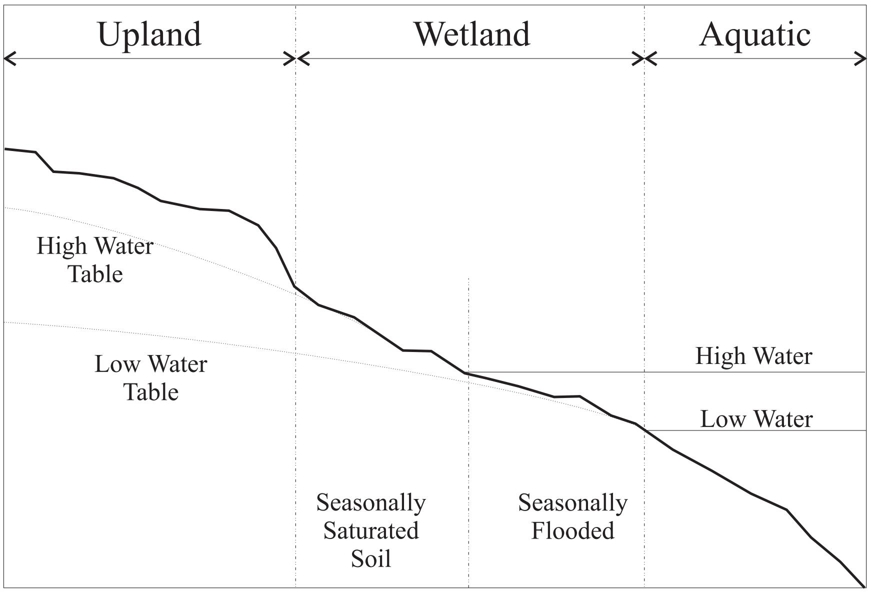


FIGURE 1.2 – Wetlands are transitional areas between between upland and aquatic systems (adapted from [2]).

transitional in the amount of water they store and process, and in other ecological processes that result from the hydrologic regime [1]. Wetland hydrologic regimes have a high variability, since they result from water flows and water storage capacities, which are very variable as well. Water enters wetlands via streamflow, runoff, groundwater discharge, tidal inflow and precipitation [2]. Wetlands lose water via streamflow, groundwater recharge, tidal outflow and evapotranspiration. The balance between wetland water storage and inflows and outflows is expressed (in general units, volume per time, [V/T]) as

$$\frac{\Delta V}{\Delta t} = P_n + S_i + G_i + T_i - ET - S_o - G_o - T_o \tag{1.1}$$

where $\frac{\Delta V}{\Delta t}$ = change in volume of water storage ΔV during a time interval Δt [V/T]
P_n = net precipitation during Δt [V/T]
S_i = surface inflows, including streamflow and runoff during Δt [V/T]
G_i = groundwater discharge during Δt [V/T]
T_i = tidal inflow during Δt [V/T]
ET = evapotranspiration during Δt [V/T]
S_o = surface outflow, including streamflow and runoff during Δt [V/T]
G_o = groundwater recharge during Δt [V/T]

T_o = tidal outflow during Δt [V/T]

The inflows and outflows are extremely variable in time, and some are stochastic. They are also variable in space, and differences between wetland water balances result in an enormous diversity in wetland types, many of them being characterized by their respective hydrologic regimes. Examples are *tidal wetlands* and *nontidal wetlands*, which can be subdivided further as *permanent wetlands* with relatively stable hydrologic conditions, *seasonal wetlands* with high seasonal water level fluctuations, or *fluctuating wetlands* with long term (several years) hydrologic fluctuations [1, 3].

1.1.2 Wetland physicochemical environment

Wetlands are transitional in terms of physicochemical setting, being sources, sinks, and transformers of nutrients and carbon [4]. Wetland physicochemistry is heavily affected by wetland hydrology (Fig. 1.1), because different sources of water inflow have different physicochemical characteristics. *Fluxial* wetlands receive nutrient and sediment rich water from upland areas by surface flow, and are generally very productive. *Pluvial* wetlands receive water exclusively from precipitation, and therefore rely on nutrients brought in from the atmosphere, generally in much lower concentrations, resulting in a low productivity, and *phreatic* wetlands receive water from groundwater inflow, and generally have a productivity status somewhere in between the former two [4]. Furthermore, the periodically or permanently submerged status of wetland soils results in a periodical decline or permanent low concentration of oxygen (oxygen diffusion is approximately 10000 times slower in submerged than in aerated soils [5]). Oxygen is the preferred oxidant in aerated soils, however, by its exclusion from wetland soils, alternative oxidants (e.g. organic substrate) must be used, thus affecting the thermodynamics and kinetics of reduction - oxidation (redox) reactions in the soil. The decline in oxygen can be measured as an increasingly negative electric potential (redox potential, Eh [mV]) and is indicative for the oxidation or reduction potential of the soil. In aerated soil (Eh > 300 mV) dissolved oxygen is prevalent, but rapidly after soil submergence the oxygen concentration and redox potential decline. A typical sequence of transformations involves: (1) nitrate reduction (Eh = 250 mV), (2) mangenese reduction (Eh = 225 mV), (3) iron reduction (Eh = [+100 -100] mV), (4) sulfate reduction (Eh = [-100 -200] mV), and (5) methanogenesis (Eh < -200 mV) [1, 2, 4]. Relative concentrations of chemical soil compounds change accordingly through time (Fig. 1.3).

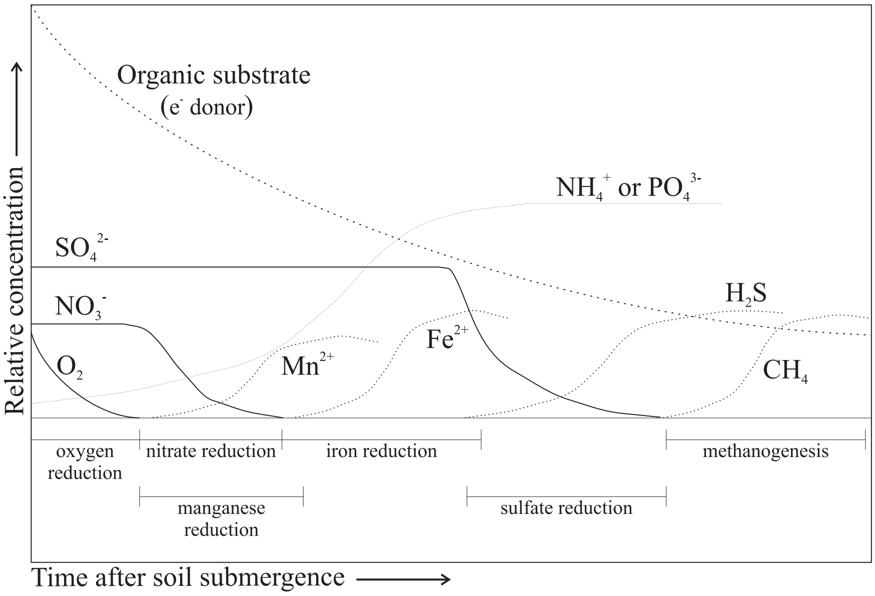


FIGURE 1.3 – Time sequence of redox transformations after soil submergence (adapted from [6]).

1.1.3 Wetland biota

Wetland environments are characterized by stresses that neither terrestrial nor aquatic organisms are adapted for to cope with [1]. Terrestrial organisms are stressed by (periodic) flooding, whereas aquatic organisms are stressed by (periodic) drought. From the wide variety of biota inhabiting wetlands, major research attention focusses on vascular plants. Soil submergence results in a variety of stresses for plants [7], of which oxygen deficiency is often the underlying factor [8, 9]. In plant cells, oxygen participates in more than 200 different reactions [10, 11]. This broad spectrum ranges from respiration, which draws on over 95% of the cellular oxygen consumption to cover the energetic needs of the cell [12], to the introduction of a double bond in a fatty acyl chain to confer the appropriate fluidity to a given membrane [13]. When plants are submitted to water-saturated soil conditions, their underground organs are facing a microenvironment that declines in oxygen concentration or even gets anoxic for a period of time. Under these conditions, the aerobic metabolism of roots of non-adapted plants shuts down and impairs the energy status of the cells, and reduces nearly all metabolically mediated activities such as cell extension and division and nutrient uptake [1, 14]. In contrast, flood-tolerant plant species (*hydrophytes*) possess a range of characteristic responses that appear to reduce the impact of the stress [15]:

- *Life history adaptations*: avoidance of adverse effects of submergence by timing of important life cycle events, such as seed dispersal, germination and reproduction;
- *Short-term metabolic adaptations*: glycolysis, ethanolic fermentation [16, 17], photosynthesis at a low CO_2 level [18];
- *Long-term responses in the root*: shift in anatomical and morphological characteristics by the formation of aerenchyma [9, 19], the formation of pneumatophores and adventitious roots [20] and radial O_2 loss to facilitate nutrient uptake [21];
- *Long-term responses in the shoot*: shoot elongation to restore contact with the open air [22, 23], potentially stimulating flowering and seed production [24].

Furthermore, wetland plants are not passive to their physicochemical and hydrologic environment (Fig 1.1), they actively affect site conditions through a variety of feedback mechanisms. Examples as peat building [25, 26], erosion reduction [27, 28], soil aeration [29], plant feedback on soil moisture [30, 31] and groundwater [32, 33] are well-documented in literature.

1.2 Wetland value and management

Wetlands provide many services and commodities to humanity [1]. Consumptive services include plant harvesting, livestock grazing, hunting and aquaculture. Non-consumptive services include recreational opportunities, water purification by retention of pollutants and sediments, flood mitigation, aquifer recharge and biodiversity conservation [1,2,34,35]. Additionally, wetland ecosystems are influencing the global cycles of water, oxygen, nitrogen, sulfur, methane and carbon dioxide at a much broader scale than the wetland itself [1, 35]. From an anthropocentric perspective, wetlands have certain *values* to the society because wetland functions have proved to be useful [36]. Wetland management is most often designed to (sustainably) exploit (some of) these in (multiple objective) management strategies. Taking the three-component conceptual model of Fig. 1.1 as a reference, it is clear that management activities can alter hydrology (e.g. ditching, draining and levee building [37, 38]), the physicochemical environment (e.g. fertilization [39]) and biota (e.g. plant harvesting [40–43]). Apart from direct wetland management, wetlands are also susceptible to indirect anthropogenic disturbances by processes such as nitrogen deposition [44, 45].

The conceptual model (Fig 1.1) is extended to account for anthropogenic disturbances (including direct wetland management and indirect disturbances). Each

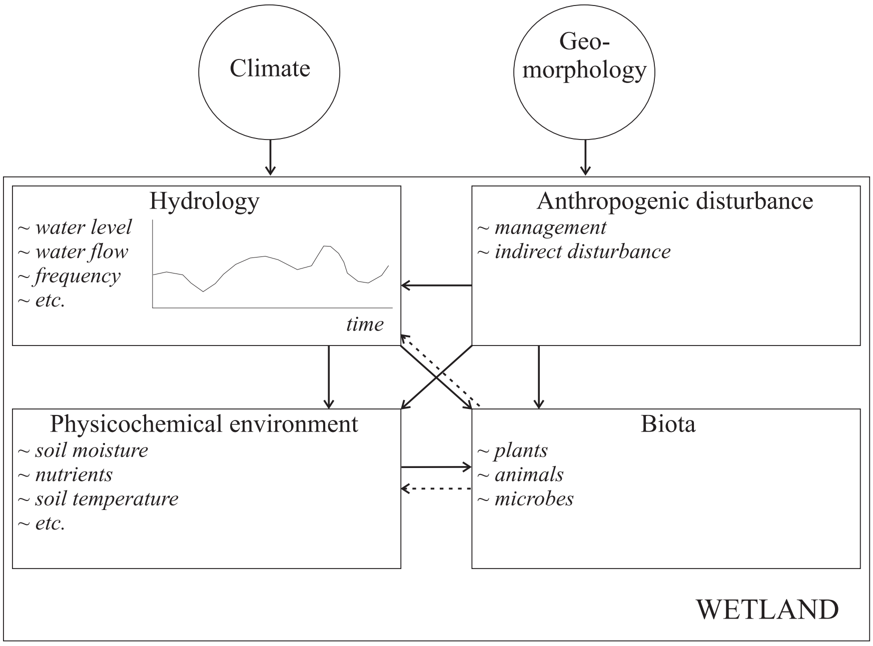


FIGURE 1.4 – Conceptual model illustrating the three wetland components (hydrology, physicochemical environment and biota) affected by a fourth component, anthropogenic disturbances.
Legend: Full arrows indicate direct effect, broken arrows feedback from biota.

of the three components of a wetland is affected by *anthropogenic disturbances* (Fig 1.4)[1].

1.3 Pattern and scale

Wetlands, as all ecological systems, exhibit *heterogeneity* on a broad range of scales [46], where heterogeneity is defined as the complexity and variability of a system property in space and time [47]. The description of complexity and variability requires the determination of *scales*, and results in the detection of *patterns*, i.e. a spatial or temporal structure that is significantly different from random, within the system [46]. Therefore the concepts of pattern and scale are closely related. Once patterns are detected, the identification of determinants and processes generating patterns results in predictive capacity [46]. The basic idea is that there

---

[1] It could be argued to include anthropogenic disturbances as a determinant, such as climate and geomorphology, or even on a higher level since anthropogenic disturbances may affect climate. In this study, however, preference is given to include anthropogenic disturbances at a lower level, and thus enhancing the workability of the wetland definition for this study.

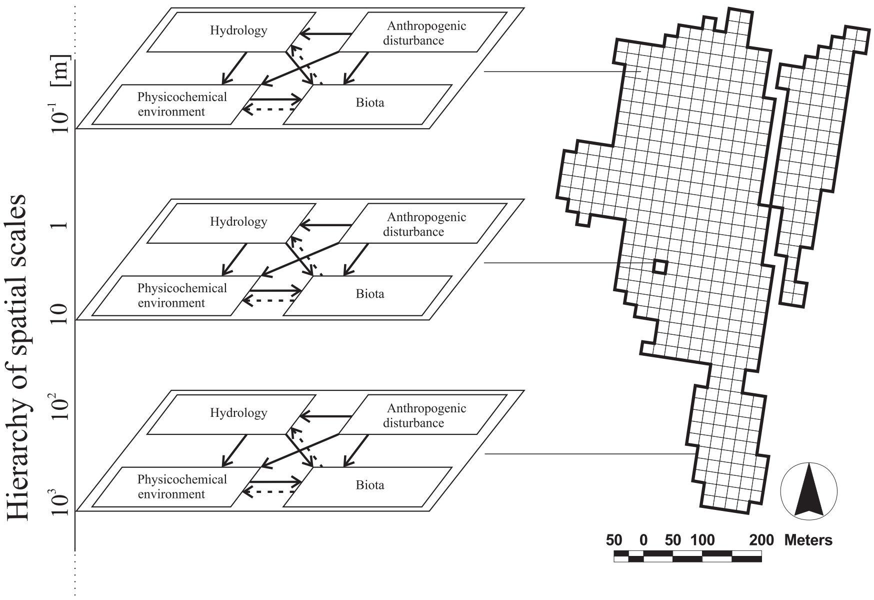

Hierarchy of spatial scales
10^{-1} [m]
1
10
10^2
10^3
Hydrology
Anthropogenic disturbance
Physicochemical environment
Biota
50 0 50 100 200 Meters

FIGURE 1.5 – Wetland heterogeneity arises from the interplay between hydrology, physicochemical environment, anthropogenic disturbances and biota acting on a hierarchy of spatial scales.

are strong bidirectional linkages between pattern and ecosystem processes [48]. But typically, patterns are observed at different scales than those at which these processes operate at [46], for example, a pattern in wetland plant species distributions may be observed at the field scale [10^2 m], while the underlying process of anaerobiosis plays at the plant root [10^{-3} m] or even root cellular [10^{-6} m] scale.

In wetlands, both spatial and temporal heterogeneity may be present in each of the four wetland components included in Fig 1.4. In fact, wetland heterogeneity arises from the interplay between hydrology, physicochemistry, anthropogenic disturbances and biota, all of them acting on a hierarchy of temporal and spatial scales [49] (Fig. 1.5). Understanding of these complex interactions, identifying the underlying driving forces and the prediction of ecosystem responses are highly relevant research topics [49], and in this context, the analysis of spatio-temporal resource distributions in relation with the distribution of vascular plant species is gaining a lot of research attention.

1.4 Patterns in the geographical distribution of species

Patterns in species distributions arise from abiotic and biotic processes acting on a hierarchy of spatial and temporal scales. The niche concept [50] facilitates interpretation. The *fundamental niche* of a plant species is a hypervolume defined by environmental dimensions in which every point corresponds to a state of the environment which would permit the species to exist and reproduce. Due to competition and other biotic interactions species generally occupy only a reduced part of this volume, the *realized niche*. The fundamental niche is primarily a function of physiological performance and ecosystem constraints, the realized niche additionally includes intraspecific and interspecific biotic interactions and competitive exclusion [51]. Interspecific differences in fundamental and realized niche result in species distribution patterns. Species distribution patterns are dynamic, showing variability in space and/or time resulting in temporarily constant spatially nonuniform patterns, or spatially constant temporarily nonuniform patterns, or spatio-temporal mosaics [46]. Therefore, species distribution patterns typically have non-equilibrium properties [52], and patterns in geographical species distributions are not static.

1.5 Ecohydrology: research at the interface between ecology and hydrology

The strong linkage between ecological and hydrological wetland characteristics resulted in the emergence of an interdisciplinary research at the interface between ecology and hydrology, *ecohydrology* (sometimes referred to as ‘hydroecology’, as a synonym or, more frequently, when the emphasis is stronger on hydrology). Hannah et al. [53] examined the evolution of the definition of ecohydrology throughout time. The first clear definition appeared in Wassen and Grootjans (1996, [54]) and covered the unidirectional nature of hydrological processes determining the natural development in wet ecosystems. Problems associated with the unidirectional nature of Wassen and Grootjans’ definition were recognized by Baird and Wilby (1999, [55]) who broadened the definition to include ecohydrological interactions of biota, mainly plant species and vegetation on hydrological processes. Additionally, Baird and Wilby argued that there is no reason why ecohydrology should be solely concerned with wetland ecosystems [55], as ecohydrological relations are important in all ecosystems. Hence, the ecohydrologic research was broadened to a range of ecosystems, including wetlands, drylands, forests, lakes, etc.

Ecohydrology investigates how hydrological processes affect plant growth and vegetation dynamics, and *vice versa* [56–59]. These ecohydrological relations can be revealed at various scale levels. Among many others, the broad scope of ecohydrological research covers the following aspects in a range of ecosystems:

- the interpretation of present soil and vegetation patterns from a hydrological point of view [60];
- the relationship between vegetation, soil and water, based on an understanding of the physiological properties of plants [55];
- the effect of hydrological regimes on vegetation succession [61];
- the development of realistic goals in ecosystem conservation [60];
- decision support in ecosystem restoration [60]; and
- the sustainable development of water resources, including socio-economic aspects [59].

Despite the recent emergence of ecohydrology as a research field at the interface between the ecological and hydrological sciences [53], its research objective, namely to understand the mutual interaction between ecosystem and hydrology, is not new [62]. Benchmarking work in the development of the ecohydrological theory is the Penman-Monteith model [63] for evapotranspiration fluxes. The model acknowledges the role of vegetation on evapotranspiration by incorporating a vegetation specific parameter, stomatal resistance (or conductance). More recently, in 2001, a collection of publications by I. Rodriguez-Iturbe, A. Porporato, F. Laio, L. Ridolfi and C. P. Fernandez-Illescas [64–67] in Advances of Water Resources presented soil moisture as the key variable for a quantitative understanding of the vegetation response to water stress. The quantitative stochastic approach presented in this work enhanced the understanding of the interplay between soil properties, climatic characteristics and vegetation water stress in savanna ecosystems. It was also inspiring to many other ecohydrological investigations in other ecosystems including wetlands, where not only root zone soil moisture dynamics are important, but also fluctuations in the water table depth [68].

1.6 Summary

The interplay between hydrology, physicochemical environment, anthropogenic disturbances and biota at a hierarchy of spatial and temporal scales define the heterogeneity of wetland ecosystems. Different tolerance levels of plant species to this heterogeneity in wetland conditions result in a pattern in plant species distributions, and hypotheses addressing (parts of) former description are typically investigated in ecohydrological research.

2

Problem definition and research objectives

2.1 Problem definition

Plant species distribution[2] in wetlands result from mutual interactions with hydrology, the physicochemical environment and anthropogenic disturbances. The study of these interactions in any wetland type at any spatio-temporal scale forms part of ecohydrological research. Within this research, a strong emphasis is on the exploration of vegetation (i.e. plant species communities) distributions in relation to the wetland environment. With the continuous development of statistical techniques, machine learning techniques and geographical information systems, modelling of these vegetation distributions based on their relation with environmental constraints have become very popular. Besides their relevance as research tools, these models are important tools to assess the impact of land use, land use changes and other environmental changes (e.g. climatic changes) on the distribution of vegetation. As such, these models can be used as management tools in wetland conservation and restoration, with a wide range of applicability.

Nevertheless, among the variety of distribution modelling techniques that have been applied in literature, efforts to introduce ensemble learning into distribution modelling remains limited until today. Ensemble learning includes techniques that

---

[2] Distributions may be random and non-random. The latter is referred to as pattern. In accordance with the use in literature, the term ‘distribution’ is used as a simplification of ‘pattern in distribution’ or ‘distribution pattern’ from this point on.

compute a collection, or ensemble, of responses, rather than a single response. This lack defines the research objectives of this dissertation.

2.2 Research objectives

The development of distribution models to predict vegetation distributions based on their relation with environmental conditions is an ongoing research task. The research objectives of this dissertation are:

- The introduction of ensemble learning by applying the so-called '*random forest*' technique in vegetation distribution modelling through the development of a *random forest distribution model* for the prediction of wetland vegetation distributions based on environmental wetland conditions;
- The assessment of the *predictive ability* of the random forest distribution model;
- The *identification of important environmental variables* determining the wetland vegetation distribution by the random forest distribution model;
- The assessment of the *generalization ability* of the random forest distribution model; and
- The analysis of input *uncertainty propagation* through the random forest distribution model.

To meet the research objectives, eight research questions should be answered:

1. Which techniques are most frequently applied for distribution modelling?
2. Can the random forest technique be used for vegetation distribution modelling?
   (a) Are there any requirements concerning data format?
   (b) Is the model output meaningful within a distribution modelling context?
   (c) Can the model output be introduced into geographical information systems?
3. Is the predictive ability of the random forest model satisfactorily?
4. Can the random forest distribution model provide information concerning the importance of environmental variables constraining the vegetation distribution? If the answer to this question is affirmative:

   (a) Would other techniques identify the same environmental variables as being important?
   (b) Is it possible to construct accurate random forest distribution models on a reduced data set, only including the most important environmental variables?

5. Is there a spatial trend in the random forest distribution modelling results?
6. Does a random forest distribution model, constructed on a given wetland, perform satisfactorily when tested on a similar but distant wetland?
7. Does the use of an ensemble modelling technique allow for uncertainty assessment?
8. How does input uncertainty propagate throughout the random forest distribution model?

2.3 Outline

Throughout this dissertation, the eight research questions are addressed and answers are given.

Chapter 1 introduced wetland ecosystems, by formulating a wetland definition in which the interplay between hydrology, physicochemical environment, anthropogenic disturbances and biota at a hierarchy of spatial and temporal scales is stressed. Based on this definition, the presence of pattern in vegetation distributions in relation to environmental conditions can be explained.

In order to model vegetation distributions based on environmental conditions, four experimental wetland sites were selected, and an abiotical and biotical characterization of these test sites is given in Chapter 3. Data from these sites are used to answer all eight research questions.

2.3.1 Research question 1

The first research question is addressed in Chapter 4 where a literature review on several statistical and machine learning techniques applied in distribution modelling highlights main properties as type and probability distribution of the response variable and prediction type of these techniques. Based on this review, the selection of the random forest technique for distribution modelling is motivated. Additionally, selection of a well-known and frequently applied technique which is used to build a reference distribution model for comparison is based on this literature review.

2.3.2 Research question 2

Chapter 4 continues by answering the second research question by constructing and calibrating a random forest distribution model for wetland vegetation based on the data provided in Chapter 3. Special attention is drawn on the input data format, the computational effort, and a meaningful spatial interpretation of the modelling results.

2.3.3 Research question 3

The model evaluation in Chapter 4 assessed the predictive ability of the random forest distribution model. Several statistical measures are used to compare observations with modelling results, and an explicit comparison is made between the random forest distribution modelling results and the results obtained by the reference model.

2.3.4 Research question 4

Research question 4 is addressed in Chapter 5 where distribution modelling is approached from two different angles: (i) predictive modelling where the predictive performance of the model is of primary concern, and (ii) explanatory modelling where the model is used to gain information on important aspects such as environmental variable importance. Under explanatory modelling, several other techniques to identify important environmental variables are included, and a comparison of results is made. Finally, results of the variable importance assessment are used for model complexity reduction. Several random forest distribution models with varying model complexity are constructed and evaluated.

2.3.5 Research questions 5–6

In chapter 6 a random forest distribution model is constructed, and results are interpreted with respect to the similarities between the different vegetation types to assess a possible spatial trend in modelling results to answer research question 5. Furthermore, the model is tested on an independent, spatially distant but ecologically similar test site to address research question 6.

2.3.6 Research questions 7–8

The random forest distribution model generates an ensemble of responses. The possibility to use such an ensemble for uncertainty assessment (research question 7) is investigated in Chapter 7 by looking into the discrete probability distribution constituted of the response ensemble. Uncertainty in distribution models originates from input data limitations, their disability to capture the entire complexity of interrelated processes resulting in vegetation distribution within the model, and the uncertainty associated with model evaluation. Among this variety of uncertainty sources, two important sources of uncertainty propagation through the random forest distribution model are selected to answer research question 8: (i) the uncertainty associated with the spatial interpolation of environmental variables, and (ii) the uncertainty associated with plant species clustering into vegetation types. The effects of this uncertainty on the random forest modelling results is investigated in Chapter 7.

3

Description of test sites: monitoring and ecohydrological review

3.1 Introduction

Four test sites are included in this study:

1. Doode Bemde;
2. Snoekengracht;
3. Vorsdonkbos-Turfputten;
4. Zwarte Beek.

All four sites are alluvial wetlands, situated in Flanders, Belgium (Fig. 3.1). They are nature reserves with relatively undisturbed abiotic and biotic conditions, with long periods of constant management (at least 10 years), and marked hydrologic gradients. The sites are included in a long term ecohydrological monitoring programme of the Research Institute for Nature and Forest (INBO), setup by W. Huybrechts and P. De Becker, and were the study sites of the Research Programme on Nature Development (projects VLINA 96/03 [70] and VLINA 00/16 [69]) of the Flemish Government. Monitoring results of the sites were gathered in four ecohydrological atlases [71–74].

The climatic conditions at the sites are typically temperate, with an average yearly rainfall of ≈800 mm distributed evenly over the year [75, 76], an aver-

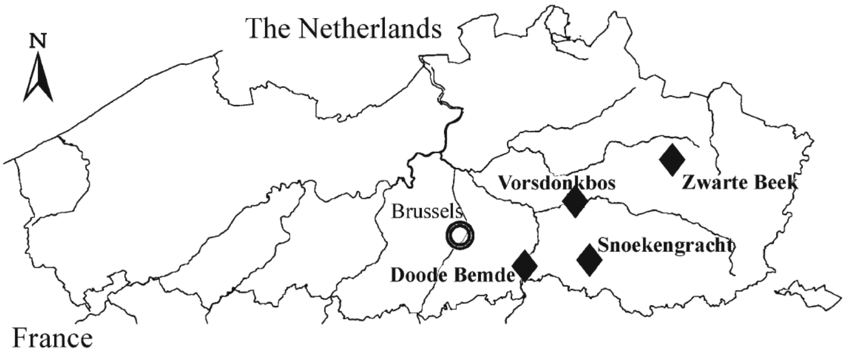


FIGURE 3.1 – Location of the test sites in Flanders, Belgium (adapted from [69]).

age annual pan evaporation of 450 mm, and an average yearly air temperature of 9.8°C [77].

Doode Bemde is an alluvial floodplain mire in the valley of the river Dijle, situated at approximately 30 m above sea level. Its soil texture is mainly loam. The area is fed by nutrient-rich groundwater (approximately 3 mm day^{-1} [78,79]). Here, a complete vegetation mosaic is found, ranging from mesotrophic alder carr and reedbeds (*Phragmitetalia*), over tall sedge swamps (*Magnocaricion*) and tall herb fen, to fen meadow and somewhat drier *Arrhenatherion elatioris* grasslands on the natural levees of the river [71, 78, 80, 81].

Snoekengracht, situated approximately 57 m above sea level, is similar to the Doode Bemde site, except for a narrower valley and even more nutrient-rich seepage water feeding the area [72, 78].

Vorsdonkbos-Turfputten is located at the southern fringe of the Demer river valley, approximately 11 m above sea level. This site is a marked seepage zone fed by two distinct aquifers. The southern part is supplied with nutrient-poor groundwater (20 mm day^{-1} [78, 79]). Here, a zone with fragments of fen grasslands (*Caricion curto–nigrae* and *Cirsio – Molinietum*) and oligotrophic woodland (*Sphagno–Alnetum*) is found. In the central and northern part of Vorsdonkbos-Turfputten, which is fed by nutrient-rich groundwater, the vegetation changes to tall herb fen (*Filipendulion*) and mesotrophic alder carr (*Caricion elongatae – Alnetum glutinosae*) [73, 78].

Zwarte Beek is situated at the western fringe of the Campinian plateau. It comprises an 800 m long section through a narrow valley, situated at approximately 52–56 m above sea level. Zwarte Beek is known for its excellent fen grasslands (mainly *Caricion curto–nigrae*). The soil consists of a 7 m thick peat layer, with an abrupt conversion into sandy sediments at the fringes of the valley. The area is fed by nutrient-poor seepage water (ca. 16 mm day^{-1} [78, 79]). The groundwater table is constant and close to the surface level throughout the year [74, 78].

3.2 Monitoring of test sites

The monitoring setup of the four test sites was very similar, both for abiotic and biotic site characterization. The sites were subdivided in regular and adjacent grid cells of 20 m × 20 m for Doode Bemde, Vorsdonkbos-Turfputten and Zwarte Beek, and 10 m × 10 m grid cells for Snoekengracht (Table 3.1), and a local coordinate system was assigned to each test site based on these subdivisions by which every monitoring location was referenced.

TABLE 3.1 – Overview of the test sites: name, abbreviation (Abbr.), location, area, grid size and number of grid cells (Nr.).

Test site	Abbr.	Location	Area	Grid size	Nr.
Doode Bemde	DB	Oud-Heverlee	20.76 ha	20 m × 20 m	519
Snoekengracht	SN	Boutersem	6.69 ha	10 m × 10 m	696
Vorsdonkbos-Turfputten	VB	Rillaar	12.80 ha	20 m × 20 m	320
Zwarte Beek	ZB	Beringen	6.80 ha	20 m × 20 m	170

3.2.1 Abiotic site characterization

Soil type was derived from hand drillings at grid cell intersections to a depth of 1 m, classified using a set of four major soil types: mineral soil with sandy texture, mineral soil with loamy texture, mineral soil with clayey texture and organic peat soil, and assigned to the neighbouring grid cells. Management focused on vegetation, and was classified per grid cell into six categories: (i) yearly mowing in early summer; (ii) cyclic mowing, once every 5–10 years; (iii) null management (no mowing or any other management regime for at least the last 10 years); (iv) transition from yearly to cyclic mowing; (v) transition from yearly mowing to no management; and (vi) transition from cyclic mowing to no management.

Piezometer networks were installed on strategic locations inside and just outside the nature reserves from 1989 onwards, and extended throughout the subsequent decade. The maximal number of piezometers differed between test sites: 36 at Doode Bemde, 36 at Snoekengracht, 40 at Vorsdonkbos-Turfputten, and 42 at Zwarte Beek. Groundwater depths [m] were measured manually every fortnight, at least during a two year period between 1991–1999. Furthermore, all piezometers were sampled on groundwater quality variables during four different sampling campaigns in spring and autumn over two consecutive years within the period 1991–1999 and included groundwater pH, K^+ [mg L^{-1}], Fe_{tot} [mg L^{-1}], Mg^{2+} [mg L^{-1}], Ca^{2+} [mg L^{-1}], SO_4^{2+} [mg L^{-1}], Cl^- [mg L^{-1}], NO_3^-–

N [mg L^{-1}], NH_4^+–N [mg L^{-1}], $H_2PO_4^-$ [mg L^{-1}] and the ionic ratio (IR = $100[1/2Ca^{2+}]/[1/2Ca^{2+} + Cl^-]$).

At Doode Bemde, samples for soil organic matter content (SOM) determination were taken at 59 locations at a depth of 0.05–0.15 m and analysed using thermal destruction at 600 °C in a muffle furnace. Soil organic matter content was expressed as a percentage [%].

3.2.2 Biotic site characterization

During spring and early summer, in the period 1993–1997, plant species occurrence (presence/absence) was mapped in the study sites on the same regular grid as soil type and management regime. Mapping was restricted to a shortlist of about 75 mainly groundwater dependent species (*phreatophytes*, [82], see Appendix A, adapted from [69]).

Species cover data were used to define vegetation types for all study sites separately using TWINSPAN [83]. Eleven clearly defined vegetation types were retained of which a short description is given in Table 3.2 and a photograph in Appendix B. All vegetation types are herbaceous, except for *Alno – Padion*, *Carici elongetae – Alnetum glutinosae* and *Sphagno – Alnetum* where a tree layer of Alnus glutinosa (L.) Gaertn. (Common Alder) among other tree species is present.

In summary, topography and piezometer locations are demonstrated in Fig. 3.2, and the spatial distribution of vegetation types is given in Fig. 4.7.

3.3 Ecohydrological review of the test sites

As stated in the introduction of this chapter, the test sites were part of two Research Programmes on Nature Development (projects VLINA 96/03 [70] and VLINA 00/16 [69]) of the Flemish Government. Furthermore, three of the test sites (Doode Bemde, Vorsdonkbos-Turfputten, Zwarte Beek) were the study areas in the dissertation of O. Batelaan [79], while I. Joris focussed in her dissertation on the Doode Bemde exclusively [84]. Based on these studies, an overview of the main properties and processes in the alluvial wetlands under investigation is given.

3.3.1 Topography

The test sites are lowland meandering river floodplains, with a characteristic topography of natural levee and lower lying flood basin (Fig 3.3). Their topography results from numerous flood deposits that create sinuous ridges along the river channels, sloping down toward the lower lying flood basin [85]. The process of lateral sediment fining results in a gradual decrease in sediment particle sizes: the

Table 3.2 – Summary of the vegetation types: number, name, short description and area.

Nr.	Name	Short description	area [ha] (number of grid cells)			
			ZB	VB	DB	SG
			6.80 (170)	12.80 (320)	20.76 (519)	6.69 (696)
1	*Alno – Padion*	Moist forest type with *Quercus robur* L., *Fraxinus excelsior* L., *Carpinus betulus* L. and some *Alnus glutinosa* (L.) Gaertn.				1.47 (147)
2	*Arrhenatherion elatioris*	High yield potential pasture, characteristic species include *Arrhenatherum elatius* (L.) J.&C.Presl., *Anthriscus sylvestris* (L.) Hoffm. and *Leucanthemum vulgare* Lamk.			2.80 (70)	0.91 (91)
3	*Calthion palustris*	Species-rich mesotrophic fen meadow dominated by species like *Caltha palustris* L., swamp horsetail *Equisetum fluvatile* L., and many *Carex*-species.			4.24 (106)	0.95 (95)
4	*Carici elongatae – Alnetum glutinosae*	Mesotrophic alder carr with dominance of *Alnus glutinosa* (L.) Gaertn. and a herblayer with *Carex acutiformis* Ehrh., *Lycopus europaeus* L. and *Solanum dulcamara* L.		3.16 (79)	1.20 (30)	1.41 (141)
5	*Caricion curto–nigrae*	Fens with small *Carex* species as *Carex panicea* L., *Carex rostrata* Stokes and *Carex nigra* (L.) Reichard.	6.80 (170)	1.12 (28)		
6	*Cirsio – Molinietum*	Comparable with *Caricion curto–nigrae* but with higher proportion of *Poaceae* and higher productivity.		1.12 (28)		
7	*Filipendulion*	Tall herb fen with *Filipendula ulmaria* (L.) Maxim., *Valeriana officinalis* L. and *Alopecurus pratensis* L.		4.76 (119)	4.16 (104)	1.12 (112)
8	*Magnocaricion*	Sedge swamp with various tall *Carex* species.			2.52 (63)	
9	*Magnocaricion with Phragmites*	*Magnocaricion* vegetation with *Phragmites australis* (Cav.) Steud.			3.72 (93)	0.83 (83)
10	*Phragmitetalia*	Highly fertile reedswamps, dominated by *Phragmites australis* (Cav.) Steud.			2.12 (53)	0.27 (27)
11	*Sphagno – Alnetum*	Oligotrophic swamp forest with *Betula pubescens* Erhr. and *Alnus glutinosae* (L.) Gaertn., with a dense moss layer of *Sphagnum palustre* L. and *Sphagnum fimbriatum* Wilson.		2.64 (66)		

ZB = Zwarte Beek, VB = Vorsdonkbos-Turfputten, DB = Doode Bemde, SG = Snoekengracht

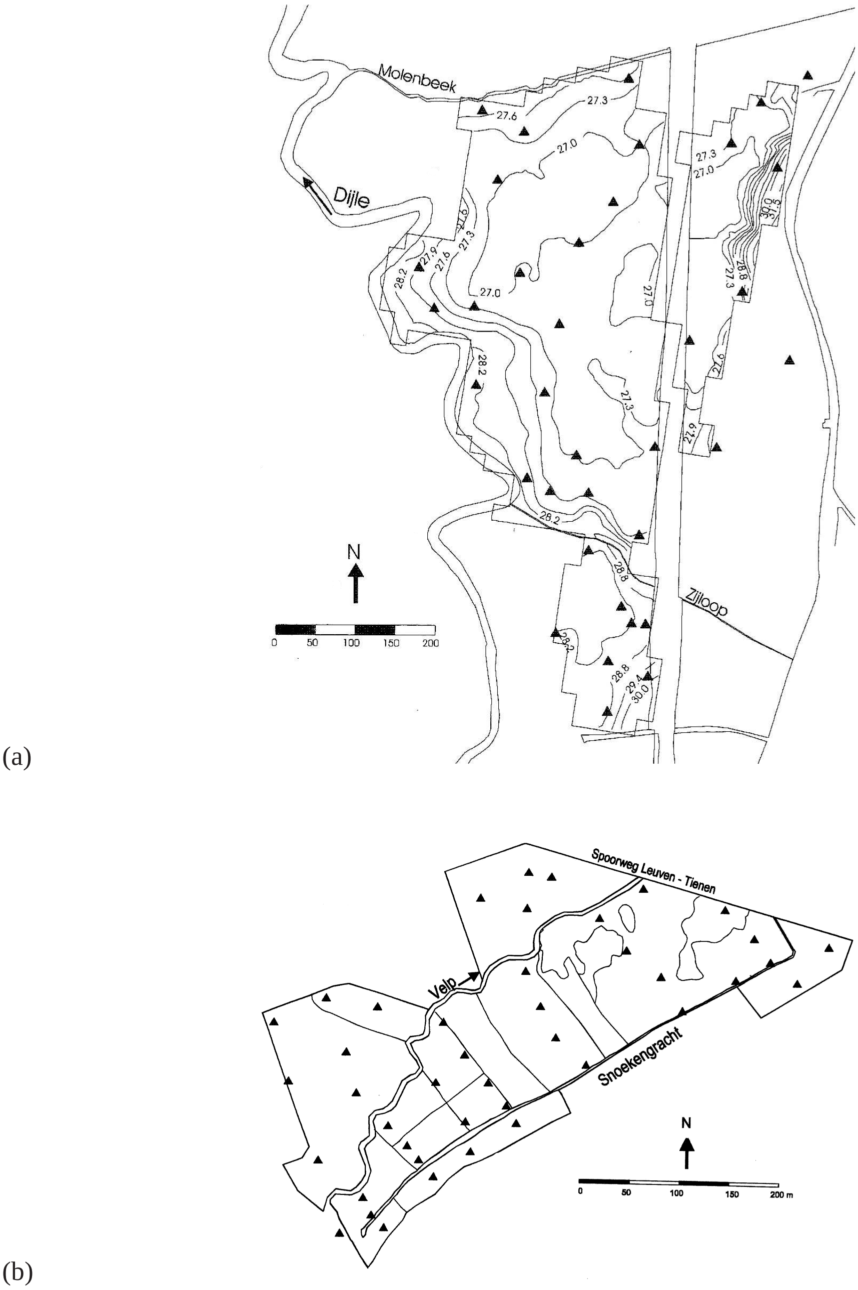


FIGURE 3.2 – Topography and piezometer (△) locations (a) Doode Bemde, (b) Snoekengracht, (c) Vorsdonkbos-Turfputten and (d) Zwarte Beek (adapted from [71–74]).

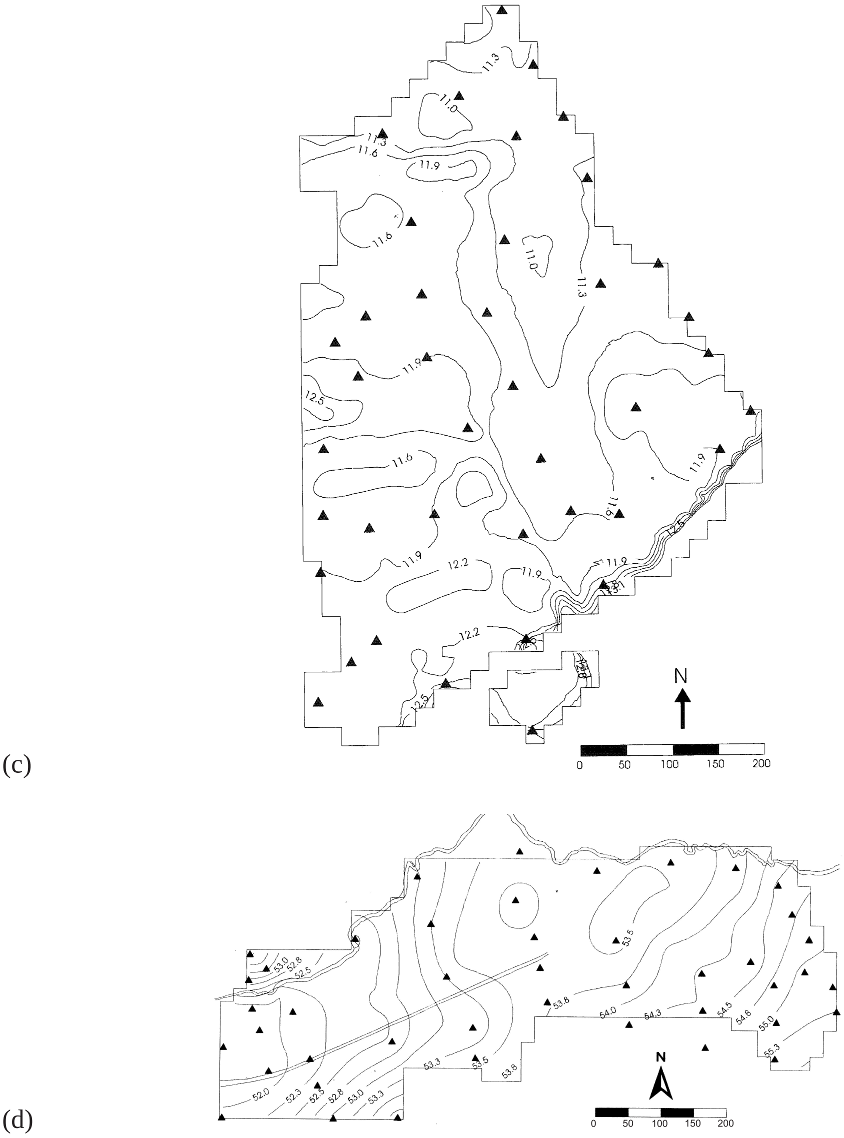


FIGURE 3.2 – *continued...*

natural levee consists of coarse sediments, while the flood basin consists of sediments with smaller particle sizes. The underlying cause is an abrupt reduction in flow velocity of the flooding water upon exiting the river channel, resulting in an immediate deposition of the coarser sediments (natural levee deposits). At distant margins of the natural levees, the deposition of loam and clay is predominant (flood basin deposits) [85]. Consequently, the topographical gradient (high levee – low flood basin) is related to the textural gradient (coarse deposits – fine deposits) in alluvial floodplains [86].

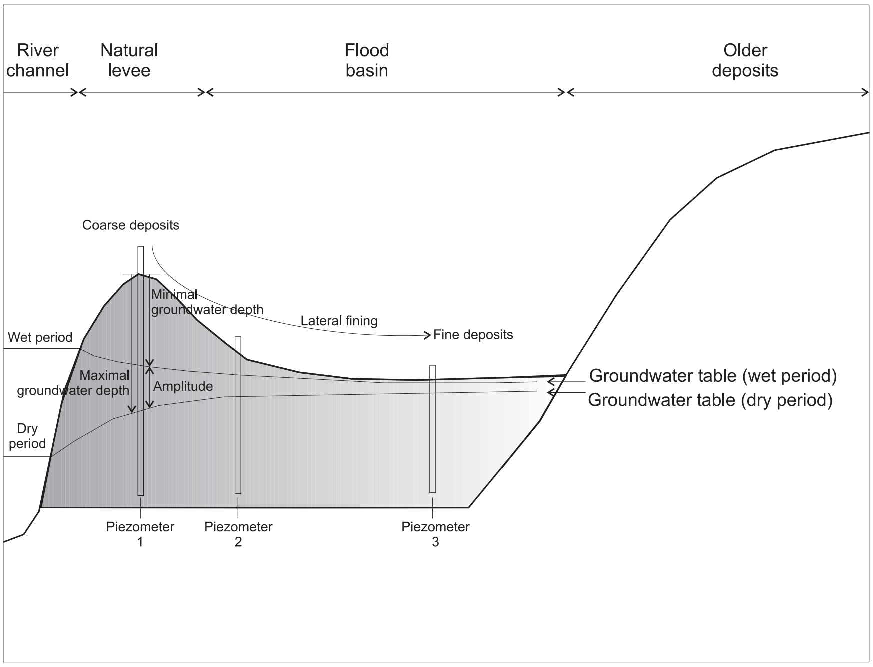


FIGURE 3.3 – Schematic cross-section of an alluvial floodplain. The gradient in grayscale intensity is illustrative for the textural gradient caused by lateral fining.

Under natural, or managed conditions allowing for overbank flooding, the levee and floodplain topography is variable in time and space, because each individual flooding event invokes spatially distributed processes of erosion, transportation and sedimentation. The hydrologic management practices at the four test sites, however, did not permit overbank flooding, and hence their topographical gradient resulting from historical flood deposits is less variable in time and space.

3.3.2 Groundwater

Hydrology is the predominant component in the wetland definition (see Section 1.1) as it directly affects numerous wetland processes. The water balance of a wetland is the total inflow subtracted by the total outflow, with in- and outflow generated by different processes (see Eq. (1.1)). Several (modelling) studies [70, 79, 84] indicated the importance of groundwater discharge and groundwater recharge (G_i and G_o in Eq. (1.1)) at the test sites, which were consequently described as *groundwater dependent wetlands* [79]. Therefore, a short description dealing with two important groundwater related aspects is given: (i) the spatio-temporal dynamics of groundwater quantity, and (ii) the groundwater quality, i.e. the hydrochemical composition of the groundwater (in terms of plant nutrients).

3.3.2.1 Groundwater quantity

Groundwater quantity is usually described by means of (consecutive) groundwater depth measurements made from piezometers. Groundwater depth measurements can be expressed relative to ground surface, or be transformed to hydrolic head by referencing to another reference level (e.g. mean sea level). Within the cross-section of an alluvial floodplain (Fig. 3.3), three piezometers are shown at different distances to the river channel: one at the natural levee (piezometer 1), another at the flood basin (piezometer 3) and a third one in between former two (piezometer 2). The depth of the groundwater table during a wet (winter) and dry (summer) period are indicated. Comparison of these depths indicates a seasonal variation: during the wet period, the groundwater depth is low (the groundwater table is close or even above ground surface), while during the dry period groundwater depth is high (the groundwater table is deep under ground surface). Therefore the groundwater depth has a time variability. Additionally, a spatial variability can be observed. At the topographically higher levee, groundwater depths are generally higher (piezometer 1), decreasing gradually toward the flood basin (piezometer 2→3) (see also further in Section 6.2).

Assume the groundwater depths in the piezometers during the wet and dry period in Fig 3.3 to be hydrologic extremes, i.e. the lowest and highest groundwater depths during a given period of time. These extremes are called the minimal groundwater depth and the maximal groundwater depth, respectively. The difference between both is the amplitude of the groundwater depth. At the study sites (as in most alluvial floodplains) a gradient of decreasing minimal groundwater depth, maximal groundwater depth and amplitude of the groundwater depth is observed from the levee toward the flood basin [71–74], and these groundwater variables characterize the groundwater quantity and dynamics.

Furthermore, alluvial systems are frequently influenced by an upward seepage flux (upward groundwater discharge), generated by a difference in hydraulic head between the recharge and discharge area. At the study sites, seepage fluxes are prevalent, ranging from approximately 3 mm day^{-1} [78, 79] at Doode Bemde to approximately 20 mm day^{-1} [78, 79] at Vorsdonkbos-Turfputten, clearly influencing groundwater dynamics [84]. In some areas within the floodplain, seepage fluxes enter the root zone of plants, while in other areas the seepage water is drained before entering the root zone. Apart from its effect on wetland water budgets and dynamics, transport of hydrochemical compounds by seepage water alters the wetland water quality.

3.3.2.2 Groundwater quality

Groundwater quality is characterized by temperature, density, specific weight, dissolved solids content, viscosity, surface tension, thermal capacity, enthalpy, vapor pressure and latent heat of vaporization [87]. The groundwater quality aspect of interest here is dissolved solids. Dissolved solids are impurities that occur because of dissolution of rocks and soils and because of the solution of CO_2 from the atmosphere [87]. Dissolved solid concentrations are usually expressed as mg L^{-1}, and natural waters contain a mixture of cation and anions including Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}, CO_3^{2-}, HCO_3^{2-}, F^{-} and NO_3^{-} in excess of 1 mg L^{-1}. Occasionally minor constituents including Fe^{2+}, Fe^{3+}, Al^{3+}, PO_4^{3-}, NH_4^{+} and NO_2^{-} achieve concentrations higher than 1 mg L^{-1} [87]. Hydrogen ions (H^{+}) are generally present in smaller concentrations (usually expressed as water $pH = -\log[H^{+}]$), but have critical influence on water chemistry by affecting dissolution and precipitation reactions.

Driven by a difference in hydraulic head, groundwater moves from higher recharge areas toward lower discharge areas. In this pathway, dissolved solutes are transported by a combination of advective and dispersive processes. The hydrochemistry of discharging groundwater in a wetland is a function of the water origin, the geochemical processes (dissolution of minerals into groundwater, precipitation of supersaturated dissolved solids, and ionization of the constituents present in groundwater) in the feeding aquifer, the travel time and the convergence of flow paths [79]. Important reversible, pH mediated net reactions are the dissolution of the carbonate minerals calcite ($CaCO_3$) and dolomite ($CaMg(CO_3)_2$):

$$CaCO_3 + CO_2 + H_2O \rightleftharpoons Ca^{2+} + 2HCO_3^{-} \tag{3.1}$$

$$CaMg(CO_3)_2 + 2H_2O + 2CO_2 \rightleftharpoons Ca^{2+} + Mg^{2+} + 4HCO_3^{-} \tag{3.2}$$

The seepage areas where groundwater containing these dissolution products discharges are higher in pH (base-rich), and gradients in base-richness have been reported to influence vegetation distributions [54].

At Doode Bemde, Vorsdonkbos-Turfputten and Zwarte Beek, different groundwater quality types of the discharging groundwater were determined [70, 79], together with a hydrochemical characterization of the precipitation and river water, sampled near the sites. Following groundwater quality types were distinguished:

(1) groundwater with a high concentration of ions, predominantly Ca^{2+} and SO_4^{2-};

(2) groundwater with a very low concentration of ions, similar to the concentrations in precipitation water, except for Ca^{2+} and HCO_3^{-};

(3) groundwater with a high concentration of ions, predominantly Ca^{2+} and HCO_3^-; and

(4) groundwater with a very high concentration of ions, predominantly Ca^{2+} and HCO_3^- but also SO_4^{2-} and Fe_{tot}.

Groundwater type 3 was the dominant groundwater type at Doode Bemde. In the each of the piezometers located within the central area of the flood basin, this groundwater type was measured. At Vorsdonkbos-Turfputten, a dominance of groundwater type 1 was observed, which extended from the entire eastern part to a significant part of the southern flood basin. Zwarte Beek was characterized by groundwater type 2, and measured groundwater concentrations were lower compared to the other two sites. For a more extensive discussion and a spatial overview of the groundwater quality at the three site, the reader is referred to the ecohydrological atlases [71, 73, 74], and to Huybrechts et al. [70] and Batelaan [79].

3.3.3 Soil solution

Quantitative (e.g. soil water content) and qualitative (e.g. nutrient concentrations) aspects of the soil solution in the root zone are important to plants (e.g. [30]). They result from the interaction between and the relative importance of different water and nutrient input and outputs. The vertically averaged soil water balance at a point in a wetland can be expressed as (modified from [30], tidal inflow and outflow neglected)

$$Z_r \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = \varphi[\theta(t),t] - \chi[\theta(t),t] \tag{3.3}$$

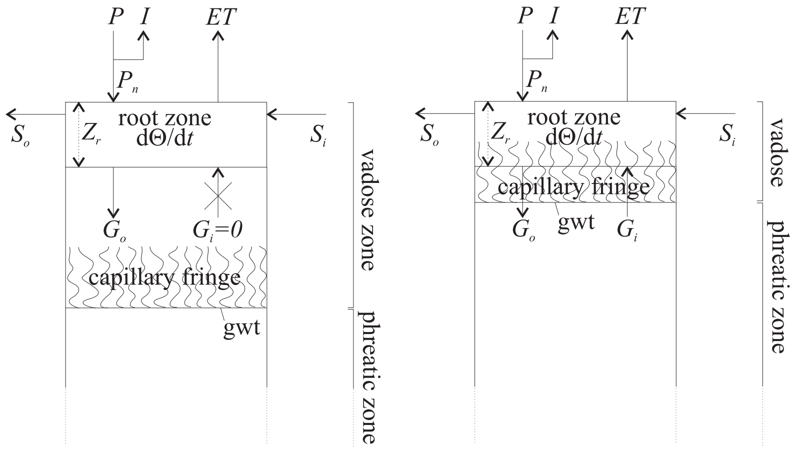
where θ [0–1 $m^3\ m^{-3}$] is volumetric soil water content ($\theta = V_w/V_s$, with the total soil volume V_s equal to the sum of the volumes of air, water (V_w) and mineral components), Z_r is the rooting depth, $\varphi[\theta(t),t]$ and $\chi[\theta(t),t]$ the soil water gains and losses from the root zone, respectively. Water reaching the soil column includes net precipitation ($P_n(t)$, i.e. rainfall minus interception), surface inflow ($S_i[\theta(t),t]$) and groundwater discharge ($G_i[\theta(t),t]$, e.g. vertical seepage fluxes and capillary rise from the water table):

$$\varphi[\theta(t),t] = P_n(t) + S_i[\theta(t),t] + G_i[\theta(t),t]\ . \tag{3.4}$$

Water is lost from the soil column by evapotranspiration ($ET[\theta(t),t]$) and groundwater recharge ($G_o[\theta(t),t]$):

$$\chi[\theta(t),t] = ET[\theta(t),t] + S_o[\theta(t),t] + G_o[\theta(t),t]\ . \tag{3.5}$$

The soil water balance is represented in Fig. 3.4. The soil column can be subdivided in two separate zones: (i) the vadose, or unsaturated, zone above the groundwater table, and (ii) the phreatic, or saturated, zone underneath the groundwater table. Water above the groundwater table is drawn upwards through continuous soil pores by capillary suction in a process called capillary rise. The soil volume in which capillary rise is present is called the capillary fringe, and its height largely depends on the soil porosity (generally smaller than 1 m in sandy soils, and up to 2 m in loamy soils [88]).



(a) (b)

FIGURE 3.4 – Water balance of a soil column where the root zone is disconnected from the groundwater table (a), and connected with the water table (b). Explanation of abbreviations: P_n net precipitation, P precipitation, I interception, ET evapotranspiration, S_i and S_o surface inflow and outflow, G_i and G_o groundwater discharge and recharge, $\frac{\mathrm{d}\theta(t)}{\mathrm{d}t}$ change in volumetric water content (θ) during a period of time t, Z_r depth of root zone, gwt groundwater table.

A spatial variability in volumetric soil water content between locations can be described to differences in soil porosity and spatial variations in the extent and relative importance of $P_n(t)$, $S_i[\theta(t),t]$, $G_i[\theta(t),t]$, $ET[\theta(t),t]$ and $G_o[\theta(t),t]$. For example, differences in soil texture and groundwater table depth between levee (Fig. 3.4(a)) and flood basin (Fig. 3.4(b)) may result in a disconnection of the soil water balance in the root zone from groundwater inputs at the former location. A temporal variability in volumetric soil water content results from the temporal stochasticity of precipitation (Eq. (3.4)) and temporal changes in the other terms of Eqs. (3.4) and (3.5). Therefore, the differences between Fig. 3.4(a) and Fig. 3.4(b) can be interpreted as a temporal difference at a given wetland location, where the former refers to a dry period and and the latter to a wet period.

Four important volumetric water contents are frequently determined: (i) the volumetric water content at soil saturation (θ_s), (ii) the volumetric water content at field capacity, i.e. 2–3 days after rainfall or irrigation (θ_{fc}), (iii) the volumetric water content at which plants wilt (θ_{wp}), and (iv) the residual volumetric soil water content (θ_r). Plant growth and vegetation development in function of these threshold values is often assessed in ecohydrological, agri- and horticultural studies (e.g. [64–67, 89]).

The nutrient concentration of the soil solution in alluvial systems is determined to a large degree by the hydrochemical composition of the water inflow, the groundwater depth and in situ soil processes [84]. The spatio-temporal variation in these forcing attributes and processes is therefore reflected in nutrient concentrations. Diurnal variations in nutrient concentrations are confined to those compounds which are strongly influenced by photosynthetic processes, e.g. dissolved oxygen [2]. Most nutrients, however, are not strongly linked nor directly linked with the solar cycle, and hence, do not display a diurnal variability. Seasonal variations are more pronounced, and related to temperature, hydroperiod, photoperiod and plant growth status [2]. In general, the growing season tends to deplete nutrients, a winter (cold temperatures, short photoperiod) and wet season tend to lower nutrient concentrations by a slow anaerobic digestion of organic matter and dilution, respectively. A dry season can accentuate organic matter decomposition and higher nutrient concentrations. It may be clear that the interactions between these driving forces is site specific and year specific.

Spatial variation in nutrient concentrations is present horizontally across the wetland area, as well as vertically within the soil column. Horizontal variation is related to spatial differences in the hydrochemical composition of the water inflows and variations in their relative importance. Spatial patterns in vegetation type and density, which are partly determined by former hydrologic wetland variability, in their turn result in spatial nutrient variations within the wetland. Over the life cycle of the plant species included in the different vegetation types, all plant tissues are either consumed, exported, or recycled back to the ground as plant litter. Wetland plant tissues fall at variable rates depending on the survival strategy of the plant species. Therefore, litterfall and subsequent nutrient release through decomposition processes are related to the spatial distribution of vegetation types. Nutrient concentrations also vary with depth in the vertical soil column. This vertical spatial variation mainly related to vertical plant root distributions and a vertical gradient in redox potential [2].

A comparison of nutrient concentrations at Doode Bemde showed differences in concentration between measurement locations and between measurement times (seasonal variation). By comparison of the hydrochemical concentration of the soil solution along a transect, perpendicular to the river Dijle, within the Doode Bemde and another transect, also perpendicular to the river Dijle, but outside the reserve on drained grounds, Joris [84] concluded that hydrochemical concentrations within Doode Bemde were similar to those of groundwater, while this similarity was not prevalent at the drained site. This horizontal spatial variation could be attributed to the seepage water flux within the wetland area, which prevents other sources of water to enter the soil column. A spatial variation in Mg^{2+} concentration was also observed, with higher concentrations measured at the natural levee, and lower concentration at the depression. A temporal variation of Ca^{2+} and HCO_3^- concentration was indicated by the same author [84]. At the natural levee, a higher biological activity within the well-aerated soils led to an increasing release of CO_2 with a subsequent increase of Ca^{2+} and HCO_3^- in the soil solution (due to an increased dissolution of calcite, see reaction 3.1). At the flood basin, the dynamics of biological activity were overruled by the hydrologic dynamics [84], and highest Ca^{2+} and HCO_3^- concentration were measured during winter when the water table was high. In summer, when the water table dropped, concentrations decreased. The reason is a higher CO_2 concentration in the soil column during winter, forcing the calcite dissolution reaction to the right. During summer, CO_2 can escape to the atmosphere, and the calcite dissolution rate lowers.

3.3.4 Soil organic matter

The organic matter of the soil arises from debris of plant material (including litterfall), animal residues and excreta that are mixed to a variable extent with the mineral soil. The dead organic matter is colonized by soil (micro-)organisms which derive energy for growth from oxidative decomposition [88]. During decomposition, essential inorganic elements are released from the organic matter in mineralization processes. There are several factors affecting the rate of organic matter decomposition, and hence soil organic matter levels. Soil properties and environmental conditions as soil water content, O_2 supply, pH and temperature are identified as key factors. The first two factors may counteract one another, when the soil water content is high O_2 gets excluded from the soil and decomposition rates lower with a possible build-up of soil organic matter. Conversely, when the soil is dry, water shortage but not O_2 will limit biological decomposition [88]. pH has only an effect by lowering decomposition rates below a pH value of 4, whereas temperature has a large effect, both by affecting plant growth, and hence by affecting litter return, and by mediating the biological activity of soil (micro-)organisms.

As can be seen in Fig. 3.5, gradients in average groundwater depth and soil organic matter content exhibit an inverse trend at Doode Bemde. At the natural levee (western border), average groundwater depths are high and soil organic matter content low, whereas the average groundwater depth at the flood basin is low and the soil organic matter content is high.

3.4 Summary

Four wetlands were selected as the study sites for this dissertation. All four are nature reserves with stable management for a considerable period of time. Within

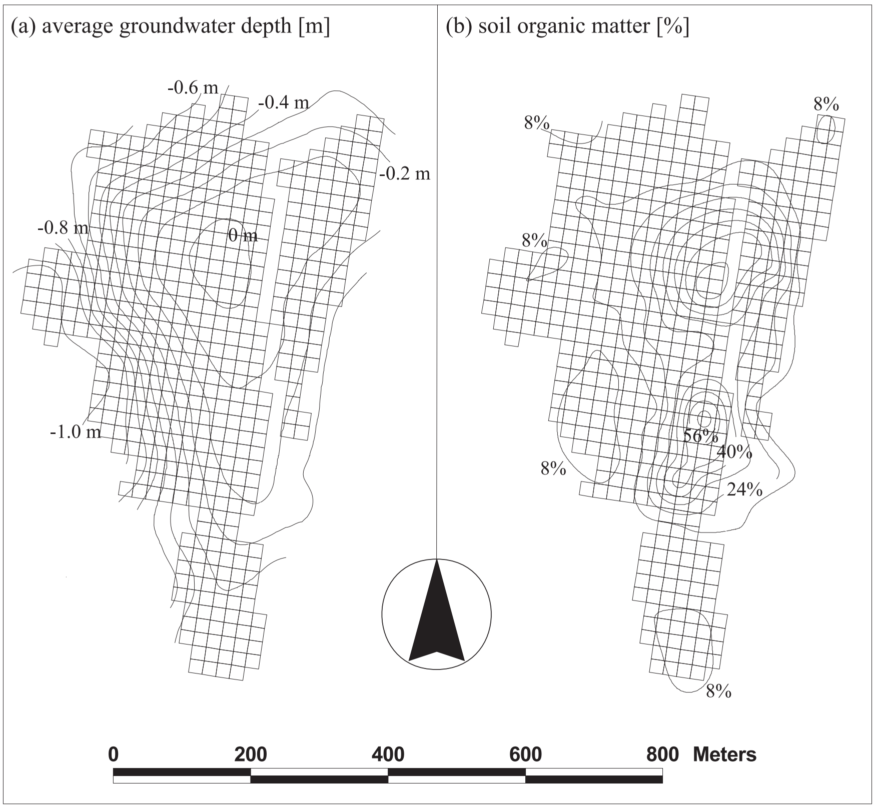


FIGURE 3.5 – Contour plot of the average groundwater depth [m below soil surface] (a) and soil organic matter content [%] (b) at Doode Bemde.

the sites, a similar monitoring protocol was set up in order to characterize the biotic and abiotic conditions. Pronounced hydrologic and hydrochemical gradients were observed, and plant species and vegetation type distributions were inventoried. Observations were gathered into ecological atlases [71–74] which serve as unique reference information on wetlands in Flanders.

4

Selection and comparison of different vegetation distribution models

4.1 Introduction

Ecohydrology tries to describe the hydrological mechanisms (like water availability and quality) that underlie ecological patterns and processes [56]. Within this scientific discipline, distribution modelling is an important issue. Several empirical models for the prediction of plant species and vegetation type occurrence in relation to hydrological or hydrogeochemical habitat conditions have been developed [90, 91]. Six empirical models, compared by Venterink [90], differ in scale level, habitat and ecosystem for which prediction was made, number of input variables, expert knowledge and field measurements requirements. However, the empirical-statistical relationship between response variable (e.g. the occurrence of species or vegetation types) and one or more explanatory variables (e.g. groundwater depth and groundwater quality variables) was generally specified by a regression model [90, 92].

Ordinary multiple regression models and multiple logistic regression models within the frameworks of generalized linear models (GLM, [93]) and generalized additive models (GAM, [94,95]) are very popular and are often used for modelling species distributions [51, 96–100]. However, other predictive distribution mod-

---

The content of this chapter is published as J. Peters, B. De Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. De Becker and W. Huybrechts. *Random forests as a tool for ecohydrological distribution modelling.* Ecological Modelling, 207:304–318, 2007.

els have been developed, based on a multitude of different modelling techniques, including neural networks (e.g. [101, 102]), ordination (e.g. canonical correspondence analysis CCA, [103]) and classification methods (e.g. classification and regression trees; [104]), Bayesian models (e.g. [105]), artificial neural networks, support vector machines, random forests, environmental envelopes (e.g. [106]) or even combinations of these modelling techniques [51].

The outline of this chapter is stipulated by the three research questions addressed (see Research questions 1–3 in Section 2.3):
1. Which techniques are most frequently used for distribution modelling?
2. Can the random forest technique be used for vegetation distribution modelling?
3. Is the predictive ability of the random forest model satisfactorily?

The most commonly used distribution modelling techniques are highlighted in a literature overview. Based on the literature overview, two different statistical techniques are further evaluated in an ecohydrological distribution modelling context: (i) multiple logistic regression and (ii) random forest. Therefore, a spatially distributed ecohydrological data set is used where 14 predictive variables, describing the abiotic environment, are related with the occurrence (presence/absence) of vegetation types. An extensive evaluation of the modelling results concludes this chapter.

4.2 Literature overview of distribution models

4.2.1 Conceptual considerations

The excellent review paper of Guisan and Zimmerman [51] highlights five core conceptual considerations for distribution modelling. In their aim to model the distribution of species or vegetation, distribution models exhibit great heterogeneity, which can be explained by these five conceptual considerations.

Conceptual consideration 1: Guisan and Zimmerman argue that Levins' classification of models [107] is useful in a conceptual context for distribution modelling [51]. Levins' model classification involves the principle that from the three desirable model properties, generality, reality and precision, only two properties can be optimized simultaneously. The third property has to be sacrificed. Hence, distribution models can be subdivided into three groups, namely distribution models that: (i) focus on generality and reality, (ii) focus on generality and precision, and (iii) focus on reality and precision. Such models are called mechanistic (or process-based), analytical (or theoretical) and empirical, respectively (Fig. 4.1).

The majority of distribution models categorize as empirical models, with a trade-off in generality. They provide a precise (including local variability) and realistic description of the ecosystem where they are constructed for, but are limited in targeting other ecosystems.

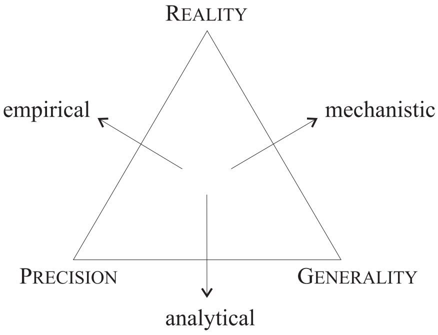


FIGURE 4.1 – A classification of models based on their intrinsic properties (Adapted from [107]).

Conceptual consideration 2: To gain model generality, it is desirable to model species or vegetation patterns based on the environmental processes (translated in model input variables) that are causal and have direct impacts on the species or vegetation pattern [51, 97]. This restrains environmental monitoring to purely scientific arguments, aiming to detect changes in time and space of environmental processes that are thought to be causal. However, monitoring efforts are frequently bounded by practical arguments such as financial issues, measurement tool limitations and site accessibility as well, which all limit the former desire.

Conceptual consideration 3: Differentiating between fundamental and realized niche [50] (see Section 1.4) distinguishes whether modelled distributions are based on theoretical physiological constraints rather than on field observations. Empirical field data include biotic interactions and therefore the distribution model will predict realized niches [51]. Distribution models based on fundamental niches should include additional rules on biotic interactions to predict realized species or community patterns.

Conceptual consideration 4: Notwithstanding the non-equilibrium state of ecosystems (see Section 1.3), most distribution models are stationary, assuming a state of equilibrium between environment and biota. For ecosystems with a lower environmental variability (e.g. permanent wetlands with relatively stable hydrologic conditions, see Subsection 1.1.1), this assumption is less restrictive, than for ecosystems with higher variability (e.g. seasonal wetlands with seasonal hydrologic fluctuations) [51].

Conceptual consideration 5: A fifth consideration comes down to two approaches having a long history in ecology: (i) clementsian discontinuous approach [108] with easily definable communities, and (ii) gleasonian continuum approach [109] with individualistic responses. Distribution models have been developed for individual species (gleasonian approach) and communities or vegetation types (clementsian approach).

4.2.2 Different distribution models

Research objectives stipulate the distribution model characteristics with regard to the previous five conceptual considerations, and hence a wide variety of distribution models have been described in literature which is too extensive to cope with in this review. Therefore, this literature review exclusively focusses on *empirical* distribution models (conceptual consideration 1) that are based on *observed data* (conceptual consideration 3) assuming an *equilibrium state* of the ecosystem (conceptual consideration 4). Such models can be conceptualized (Fig. 4.2) reverting to the wetland definition from Section 1.3.

The distribution model can be specified by various techniques. The most popular statistical techniques together with machine learning techniques, only recently applied in distribution modelling, are highlighted.

Generalized linear models

A generalized linear model (GLM) provides a way of estimating a function of the mean response (the *link function*, $g(\mu)$) as a linear combination of some set of predictive variables [93, 110, 111]:

$$g(E(Y|\mathbf{x})) = g(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \eta(\mathbf{x}) \tag{4.1}$$

where $E(Y|\mathbf{x})$ is the expected response given $\mathbf{x}$, g is the link function and $\eta(\mathbf{x})$ the linear predictor, a linear function of the predictive variables x_i with parameters β_i. Depending on the distribution of the response variable, different link functions are used (Table 4.1).

The GLM most frequently applied in distribution modelling is a GLM setup with the logit link function, the so-called *logistic regression* models. In logistic regression models, a binary response (taking only values 0 and 1) is modelled by a linear combination of predictive variables. Binary responses (i.e. presence/absence of species) are prevalent in ecology. Logistic regression model predictions take values of 0 (absence), 1 (presence) and all values in between these extremes. Predicted values are interpreted as the probability of occurrence of a species, community or vegetation type in distribution modelling.

If the response is not linear with any of the predictive variables x_i included in $\eta(\mathbf{x})$, a transformed term of x_i can be included in the model [51]. Several trans-

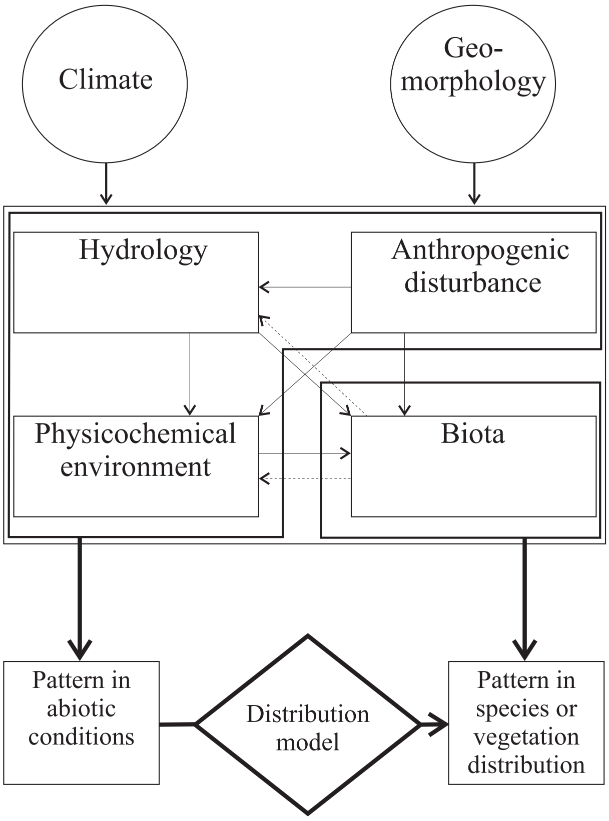

Climate
Geo-morphology
Hydrology
Anthropogenic disturbance
Physicochemical environment
Biota
Pattern in abiotic conditions
Distribution model
Pattern in species or vegetation distribution

FIGURE 4.2 – Conceptualization of distribution models. The pattern in abiotic wetland conditions are used to model species and vegetation distributions assuming a state of equilibrium.

formation functions with a strict parametric functional form [112] have been used including second order polynomials [113, 114], third order polynomials [113], β-functions [115, 116], a hierarchical series of models (HOF model, [117]) and a set of *n*-transformed functions [118].

Since their development by Nelder and Wedderburn in 1972 [110], GLMs have been included in numerous ecological and ecohydrological studies. An overview of GLM applications in distribution modelling is given in Table 4.2.

Generalized additive models

In GLM the predictor is a linear function of the model parameters. The generalized additive model (GAM), which has been developed by Hastie and Tibshirani in 1990 [94], extends the generalized linear model by fitting nonparametric smoothing functions to estimate relationships between the response and the predictive

TABLE 4.1 – Exampes of link functions for several distributions.

Distribution	link function (g)
Gaussian distribution	identity link function
Exponential distribution	inverse link function
Gamma distribution	inverse link function
Inverse Gaussian distribution	inverse squared link function
Poisson distribution	log link function
Binomial distribution	logit link function
Multinomial distribution	logit link function

variables. The form of a GAM is:

$$g(E(Y|\mathbf{x})) = g(\mu) = \alpha + \sum_{i=1}^{p} f_i x_i = \eta(\mathbf{x}) \tag{4.2}$$

where g is the link function, α a constant intercept, and f_i the nonparametric function describing the relationship between the transformed mean response $g(\mu)$ and the i-th predictive variable x_i. $\eta(\mathbf{x})$ is referred to as the additive predictor. The nonparametric functions are estimated from the data using smoothing operations, and they include running means, locally weighted regression, or locally weighted density functions [51]. An overview of different GAMs and case studies found in literature are given in Table 4.2.

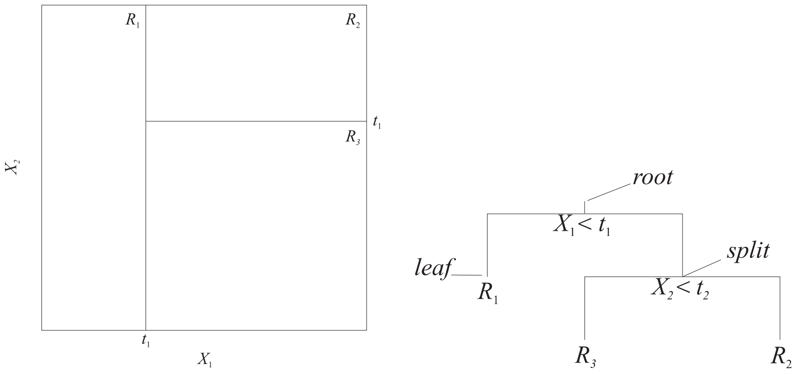
Tree-based techniques

Tree-based techniques partition the predictor space (here, environmental space) into parts, and then fit a simple model (like a constant) to each part [139]. Classification (categorical response) and regression (continuous response) trees (CART, developed by Breiman et al., 1984 [104]) is a popular technique, and other techniques such as rule-based classification [140] and maximum likelihood classification [141] have been developed.

CART uses recursive binary partitioning to split the predictor space. In a first step, the predictor space is split into two regions choosing a predictive variable (X_i) and cutpoint (t_i) to achieve the best fit. Then one or both regions are split into two more regions, and this process is iterated until a stopping rule is reached [139]. For example, a two-dimensional predictor space (predictive variables X_1 and X_2) is first split at $X_1 = t_1$, resulting in two regions $X_1 < t_1$ and $X_1 \geq t_1$. Then $X_1 \geq t_1$ is split into $X_2 < t_2$ and $X_2 \geq t_2$, resulting in three regions R_1, R_2 and R_3 (Fig. 4.3(a)). The corresponding tree is shown in Fig. 4.3(b). Starting with data at the top node of the tree (*root*), a rule for creating new branches (*splitting rule*) recursively splits the data in each *node* until a stopping rule is reached, and a terminal node of the tree (*leaf*) is reached. Then a model is fitted to the leaf (e.g. leaf average in regression

TABLE 4.2 – Overview of generalized linear models (GLM) and generalized additive models (GAM) in distribution modelling with information on the response variable (type, probability distribution and examples), the model prediction and literature examples. No case studies found is indicated by —.

Modelling technique	Response variable			Prediction	Case studies in literature
	Type	Probability distribution	Example		
GLM	continuous	Gaussian	percent cover species richness biomass	probability	[119–121]
GLM	continuous	Poisson	individual counts species richness	probability	[122–124]
GLM	continuous	Negative binomial	individual counts	probability	[122]
GLM	categorical	Binomial	presence/absence relative abundance	probability	[78, 97, 98, 100, 114, 125–131]
GLM	categorical	Multinomial	vegetation type plant community	probability	[96]
GAM	continuous	Gaussian	percent cover species richness biomass	probability	—
GAM	continuous	Poisson	individual counts species richness	probability	[132]
GAM	continuous	Negative binomial	individual counts	probability	—
GAM	categorical	Binomial	relative abundance relative abundance	probability	[78, 92, 95, 125, 133–138]
GAM	categorical	Multinomial	vegetation type plant community	probability	—



(a) (b)

FIGURE 4.3 – Recursive binary partitioning using CART. Partition of a two-dimensional predictor space (a) and the corresponding tree (b).

trees, leaf majority vote in classification trees, among others [142]). A selection of tree-based distribution models in literature are given in Table 4.3.

TABLE 4.3 – Overview of distribution models using tree-based techniques, canonical correspondence analysis (CCA), redundancy analysis (RDA), Bayesian techniques, artificial neural networks (ANN), support vector machines (SVM), and random forest (RF).

Modelling technique	Response variable	Prediction	Case studies
Regression tree	continuous	response value	[143–150]
Classification tree	categorical	class	[113, 125, 130, 135, 145, 146, 150–161]
CCA	continuous	distribution	[127, 162–171]
RDA	continuous	distribution	[162, 172–174]
Bayesian	continuous	probability	[175]
Bayesian	categorical	probability	[176–178]
ANN	continuous	probability	[125, 175, 179–182]
ANN	categorical	probability	[125, 183–185]
SVM	categorical	class	[186–188]
RF	categorical	probability	[131, 189–191]

Canonical ordination techniques

Canonical ordination techniques are designed to detect patterns in biotic variation that can be explained best by the observed environmental variables [192]. Canonical ordination includes canonical correspondence analysis (CCA, [193]) developed in the mid-eighties, redundancy analysis (RDA, [194, 195]) developed by

Rao in 1964, canonical correlation analysis [196,197] and canonical variate analysis [197,198]. Most distribution models based on ordination techniques, however, use CCA [51]. CCA relates biotic variation (e.g. variation in species or vegetation occurrence) directly to environmental variation by combining a multivariate ordination of biotic data with a constrained regression maximizing the correlation between the ordination axes and selected environmental variables [97]. The biota are assumed to have unimodal responses, with equal widths (tolerances) and amplitudes (maxima) to the underlying environmental gradients as specified by the ordination axes, with modes distributed uniformly along an environmental gradient that is long compared with species tolerances [51, 97, 193, 199].

RDA assumes linear distributions along environmental gradients, which constrains the applicability of this technique to short environmental gradients and limits its use in distribution modelling. Examples of distribution models using CCA and RDA are given in Table 4.3.

Bayesian techniques
Distribution models based on Bayes' theorem modify an initial (*a priori*) estimate of the probability of encountering a species or vegetation type in the landscape by using the known preferences of a species or vegetation type for environmental characteristics and information concerning the distribution of these characteristics in the landscape [51, 176]. These models effectively use *a priori* knowledge (e.g. expert knowledge, literature knowledge) to an *a posteriori* prediction of occurrence given known environmental characteristics. The inclusion of *a priori* informations in these models may be advantageous resulting in a more efficient scientific process. However, using inaccurate, invalid or unappropriate *a priori* information decreases Bayesian model performance drastically [200]. Examples of distribution models using Bayesian techniques are given in Table 4.3.

Artificial neural networks
Artificial neural networks (ANN) encompass a group of learning algorithms that linearly combine predictive variables to model a response as a nonlinear function of these predictive variables [139,201]. They are designed after the functioning of the human brain, and consist of an input layer, a (number of) hidden layers, and an output layer. Artificial neural networks apply both to regression (continuous response) and classification (categorical response) [139] (Table 4.3) and are useful in distribution models when underlaying data relationships are unknown [201].

Support vector machines
Support vector machines (SVM, [202, 203]) are machine learning techniques that aim to separate the predictor space maximally. Linear support vector machines achieve this by calculating optimal separating hyperplanes maximizing the margin between the elements from the different classes. Support vector machines can be used for classification and regression, but for distribution modelling only classi-

fication examples were found (Table 4.3). Support vector machines and artificial neural networks were developed by the end of last century, and their application into distribution modelling studies is only recent.

Random forest
Random forest (RF, developed by Breiman, 2001 [204]) is an ensemble learning technique which generates many unpruned classification (categorical response) or regression trees (continuous response), that are aggregated to compute the final result. The main differences with ordinary tree-based techniques (as described earlier) lies in the generation of an ensemble of trees by the random forest technique, whereas ordinary tree-besed techniques tend to construct the single best tree. The construction of an ensemble of tree-based classifiers in random forest, however, results in a higher accuracy compared with individual tree-based classifiers. The application of random forest in distribution modelling is restricted to classification studies so far (Table 4.3).

Other techniques
The list of techniques described before is not exhaustive. A multitude of other distribution modelling techniques exist: environmental envelope (e.g. included in BIOCLIM [205, 206]) and HABITAT [207] model, distance based techniques included in DOMAIN [208] and LIVES [209] model, matrix regression combined to generalized linear modelling within a generalized dissimilarity modelling context [209], multivariate adaptive regression splines, maximum entropy models and boosted regression trees [209].

4.2.3 Selection of techniques

Two of the techniques shortly described above, were selected for evaluation in this chapter: (i) multiple logistic regression, within the framework of GLM (i.e. a GLM using the logit link function), and (ii) random forest. The choice for the multiple logistic regression technique was based on the nature of the response variable (binomial for each individual vegetation type) and its extensive use in distribution modelling studies on various ecosystems, including the test sites presented in Chapter 3 [70, 78]. The second technique selected was random forest, a recently developed ensemble learning technique. The use of this technique proved to be successful in several scientific areas, however, its application in distribution modelling was not tested. The novelty of the study presented in this chapter is the introduction of random forest in distribution modelling. In fact, the multiple logistic regression was used as a reference technique to evaluate the random forest technique against.

4.3 Material and methods

4.3.1 Study sites

The study sites included Doode Bemde, Snoekengracht, Zwarte Beek and Vorsdonkbos-Turfputten, covering an area of 47.05 ha, subdivided in 1705 grid cells (see Table 3.1). The abiotic and biotic site characterization of the test sites is described in Chapter 3. Abiotical monitoring included groundwater depth, several groundwater quality variables, soil type and management. Biotical site characterization resulted in a spatially distributed vegetation map, as demonstrated in Fig. 3.2.

4.3.2 The ecohydrological data set

For each of the four sites, an observation data set was constructed, including the average groundwater depth (AGD, [m]) derived from piezometer groundwater depth measurements, the groundwater quality variables, including groundwater pH, K^+ [mg L^{-1}], $Fe_{(tot)}$ [mg L^{-1}], Mg^{2+} [mg L^{-1}], Ca^{2+} [mg L^{-1}], SO_4^{2+} [mg L^{-1}], Cl^- [mg L^{-1}], NO_3^-–N [mg L^{-1}], NH_4^+–N [mg L^{-1}], $H_2PO_4^-$ [mg L^{-1}] and the ionic ratio (IR = 100[1/2Ca^{2+}]/[1/2Ca^{2+} + Cl^-]) measured from piezometer groundwater samples, and the spatial coordinates of the piezometer locations. A spatial interpolation using block kriging (for details, see [70, 78]) was conducted in order to obtain groundwater variable estimates for all 1705 grid cells. Together with the other abiotic (soil type and management regime) and biotic (vegetation type) variables, groundwater variables were transferred to a data set. The data set contains 1705 measurement vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i14})$ constituted of the values of 14 predictive variables (12 continuous and 2 categorical), describing the abiotic environmental conditions. Eleven different vegetation types $c_1, \ldots, c_{11}$ are considered (Table 3.2). To each measurement vector $\mathbf{x}_i$ a unique vegetation type l_i is assigned. This data set will be referred to as 'ecohydrological data set' and is denoted as (N = 1705):

$$L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\} . \tag{4.3}$$

4.3.3 Statistical model description

4.3.3.1 Multiple logistic regression

Multiple logistic regression describes the relationship between a combination of environmental variables and a binary response variable by means of a link function, the logit transformation [210]. Consider a collection of p independent predictive variables denoted by the vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, and let the conditional probability that the outcome is 'present' be denoted by $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, then the logit transformation (logit$[\pi(\mathbf{x})]$) is used to relate the independent predictive variables with a binomial (0/1) distributed response. The logit link function $g(\mathbf{x})$ is given by [210]:

$$g(\mathbf{x}) = \text{logit}[\pi(\mathbf{x})] = \ln\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\,, \qquad (4.4)$$

and the multiple logistic regression model is:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}\,. \qquad (4.5)$$

Eq. (4.5) results in a sigmoid curve with a low/high probability that the outcome is present for a big range of low/high $g(\mathbf{x})$ values, and a steep increase in probability in the middle section of the plot (Fig. 4.4).

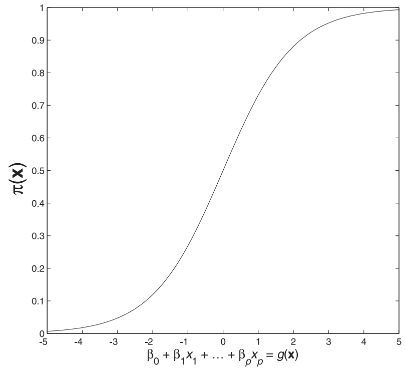


FIGURE 4.4 – The relationship between a linear combination of the independent variables and the conditional probability that the outcome is present as response variable.

If some of the predictive variables are categorical (e.g. soil type and management in the ecohydrological data set), it is inappropriate to include them in the model as such. In that case a collection of design variables (or dummy variables) is to be used. In general, if a categorical predictive variables has k possible values, $k-1$ design variables are needed. When, for example, the j–th predictive variable is soil type with four possible classes, i.e. sand, loam, clay or peat, three design variables are necessary. A possible coding strategy makes use of Helmert contrasts [211] (Table 4.4).

The link function for multiple logistic regression with p environmental variables and the j–th predictive variable being categorical would be

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \left(\sum_{l=1}^{k_j-1} \beta_{jl} D_{jl}\right) + \ldots + \beta_p x_p = \text{logit}[\pi(\mathbf{x})] \qquad (4.6)$$

TABLE 4.4 – Translation of a categorical predictive variable into design variables using Helmert contrasts.

			Design variables				
			D_1	D_2	D_3	$\cdots$	D_{k-1}
Categorical pred.	k classes	1	-1	-1	-1		-1
		2	1	-1	-1		-1
		3	0	2	-1		-1
		4	0	0	3		-1
		$\vdots$					
		k	0	0	0		$k-1$

where D_{jl} are the values of $k_j - 1$ design variables.

An estimator $\hat{g}(\mathbf{x})$ for the logit function has to be found for each response class (vegetation type) separately, in order to get an estimation of the probability of occurrence, $\hat{\pi}(\mathbf{x})$, according to Eq. (4.5). The estimation is based on maximization of the likelihood function, yielding values for the unknown parameters $\beta = (\beta_0, \ldots, \beta_p)$ which maximize the probability of obtaining the observed set of data under the multiple logistic regression model. Hosmer and Lemeshow (2000) [210] derive and express the likelihood function as follows. $P(Y = 1|\mathbf{x})$ denotes the conditional probability that $Y = 1$ given $\mathbf{x}$, which can be determined using Eq. (4.5) and equals $\pi(\mathbf{x})$. Therefore, the quantity $1 - \pi(\mathbf{x})$ expresses the conditional probability that Y is equal to 0 given $\mathbf{x}$, $P(Y = 0|\mathbf{x})$. For those pairs $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i \in X$ and $y_i \in Y$, where $y_i = 1$, the contribution to the likelihood function is $\pi(\mathbf{x}_i)$, and for those pairs where where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(\mathbf{x}_i)$. The likelihood function for pair $(\mathbf{x}_i, y_i)$ is expressed as

$$\pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} \,. \tag{4.7}$$

The likelihood function ($l(\beta)$) is obtained as the product of the terms given in Eq. (4.7) over all N pairs:

$$l(\beta) = \prod_{i=1}^{N} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} \,. \tag{4.8}$$

The estimates for $\beta = (\beta_0, \ldots, \beta_p)$ maximize the expression in Eq. (4.8). However, the likelihood function is most commonly expressed as the log-likelihood

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{N} y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)] \,. \tag{4.9}$$

for mathematical convenience.

A full model, including first order terms and quadratic variable terms (not included in Eqs. (4.4) and (4.6)), was fitted to the data using the likelihood function. Afterwards, stepwise insertion or deletion of variables [210] was applied. A bi-directional stepwise model selection procedure was used, starting with the full model and alternately omitting and re-introducing one model component at each step. Selection stopped when no predictive variable insertion or deletion caused a lower Akaike Information Criterion value (AIC, [212]), resulting in the model with the lowest AIC value. Akaike's Information Criterion calculates a trade-off value between the model goodness-of-fit and the model complexity:

$$\mathrm{AIC} = 2 \times k - 2L(\beta)$$

with k the number of estimated model parameters and $L(\beta)$ the log-likelihood function.

Model goodness-of-fit is assessed by the deviance (D), likelihood ratio (G) and Pearson chi-square statistics. The deviance statistic compares observed values of the response variable with predicted values obtained from the model. Hosmer and Lemeshow [210] propose to think of an observed response value as a realization of a fully saturated logistic regression model, i.e. a model with as many parameters as there are data points. The comparison of observed to predicted values is based on the likelihood function:

$$D = -2\ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right]. \tag{4.10}$$

Substituting Eq. (4.9) into Eq. (4.10) results in

$$D = -2\sum_{i=1}^{N} y_i \ln\left[\frac{\hat{\pi}(\mathbf{x}_i)}{y_i}\right] + (1-y_i)\ln\left[\frac{1-\hat{\pi}(\mathbf{x}_i)}{1-y_i}\right]. \tag{4.11}$$

When the response variable has a binary distribution with values 0 and 1, as is the case in this study, the likelihood of the saturated model equals 1, simplifying the deviance statistic to

$$D = -2\ln(\text{likelihood of the fitted model}). \tag{4.12}$$

The smaller the difference between the log-likelihood of the fitted and saturated model, the smaller is the deviance and the closer is the fitted model to the perfectly fitting, saturated model. A larger deviance indicates a poorer model fit. The deviance of null models (intercept only) is given by D_{null}, the deviance of fitted models by residual deviances (D_{resid}). If D_{resid} is smaller than the corresponding chi-square value ($\chi^2(1-\alpha, \mathrm{Df})$), the logistic regression model is concluded to be appropriate, providing an adequate fit.

The likelihood ratio (G) is used when the significance of an independent variable is assessed. G compares the deviance of a model without and a model with

TABLE 4.5 – Instances with equal measurement vectors are grouped to determine Pearson residuals and Pearson goodness-of-fit.

number of groups	1	2	3	4	…	J
number of positive responses	f_1	f_2	f_3	f_4	…	f_J
predicted response	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	…	$\hat{\pi}_J$
number of group members	m_1	m_2	m_3	m_4	…	m_J

the independent variable of interest as:

$$G = D(\text{model without the variable}) - D(\text{model with the variable}) \tag{4.13}$$

or the null model (intercept only) with the residual model with parameters $\beta_1, \beta_2, \ldots, \beta_p$, to test the null hypothesis $H_0 : \beta_i = 0\ (i = 1, \ldots, p)$. G follows approximately a chi-square distribution with degree of freedom corresponding to the difference in degree of freedom for the two models in comparison.

The Pearson chi-square goodness-of-fit equals the sum-of-squares of the Pearson residuals. The calculation of Pearson residuals starts from grouping instances (i.e. grid cells) with equal measurement vectors $\mathbf{x}$. If J denotes the number of distinct measurement vectors within the observed data set containing N instances, than $J < N$ if some instances have equal measurement vectors. Grouping instances having equal measurement vectors together results in J groups, where each group j consist of m_j instances ($j = 1, 2, 3, \ldots, J$). It holds that $\sum m_j = N$. The observed number of positive responses ($y_j = 1$) among the m_j instances is denoted by f_j. An overview of the grouping procedure is given in Table 4.5. The estimated number of positive responses for group j ($j = 1, 2, 3, \ldots, J$), $\hat{f}_j$, is

$$\hat{f}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1 + e^{\hat{g}(\mathbf{x}_j)}}, \tag{4.14}$$

where $\hat{g}(\mathbf{x}_j)$ is the estimated logit. The Pearson residual for a particular group is defined as follows:

$$r(f_j, \hat{\pi}_j) = \frac{(f_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}. \tag{4.15}$$

The sum-of-squares of these residuals is the Pearson chi-square statistic

$$X_p^2 = \sum_{j=1}^{J} r(f_j, \hat{\pi}_j)^2. \tag{4.16}$$

Pearson residuals can be used to identify outlying instances. In this study, where 12 continuous predictive variables are used, a reasonable expectation is that $J \approx N$. Therefore, Pearson residuals are calculated for (almost) every instance. Instances with high residual values indicate a pronounced difference between observed and expected response, and can therefore be eliminated from the data set.

The assumption that the Pearson goodness-of-fit statistic has a chi-square distribution with $J-(p+1)$ degrees of freedom, however, is incorrect when $J \approx N$. To avoid this problem, a limited number of groups has to be defined (e.g. number of groups = ceiling($2 \times N^{2/5}$), [213]), for which observed and expected frequencies are calculated.

4.3.3.2 Random forests

The random forest technique [204] is an ensemble learning technique which generates many classification trees [104] that are aggregated to compute a classification. A necessary and sufficient condition for an ensemble of classification trees to be more accurate than any of its individual members, is that the members of the ensemble perform better than random and are diverse [214]. Random forests increase diversity among the classification trees by resampling the data with replacement, and by randomly changing the predictive variables sets over the different tree induction processes. Each classification tree is grown using another bootstrap subset X_i of the original data set X and the nodes are split using the best split variable among a subset of m randomly selected predictive variables [215]. This is in contrast with standard classification tree building, where each node is split using the best split among all predictive variables. The number of trees (k) and the number of variables to split the nodes (m) are two user-defined parameters. The number of trees (k) equals the number of bootstrap subsets used to construct the random forest, since one classification tree is constructed based on one bootstrap subset. Predictive variables may be continuous or categorical, circumventing the need to translate the latter into design variables. The algorithm for growing a random forest of k classification trees is given in Algorithm 1. Additionally, an unbiased estimate of the generalization error (the so called out-of-bag error, oob error [204]) is obtained during the construction of a random forest (Algorithm 2).

Breiman [204] proved that random forests produce a limiting value of the generalization error. As the number of trees increases, the generalization error always converges. The number of trees (k) needs to be set sufficiently high to allow for this convergence. Consequently random forests do not overfit. An upper bound of the generalization error can be derived in terms of two parameters that measure how accurate the individual classification trees are and how diverse different classification trees are [204]: (i) the *strength* of each individual tree in the forest; and (ii) the *correlation* between any two trees in the forest. A classification tree with a low error is a strong classifier. Strength and correlation are not user-defined parameters. However, reducing the number of randomly selected predictive variables to split the nodes (m) decreases both strength and correlation. Decreasing the strength of the individual trees increases the forest error. Whereas decreasing the correlation decreases the forest error. Therefore m, which is a user-defined parameter, has to be optimized in order to get a minimal random forest error.

Algorithm 1: The construction of a random forest.

```
Data: training data set $X$
Result: random forest consisting of $k$ classifiers
define parameters $m$ and $k$;
for $i = 1$ to $k$ do
    draw a bootstrap subset $X_i$ containing approximately 2/3 of the elements of the original data set $X$;
    use $X_i$ to grow an unpruned classification tree to the maximum depth, with the following modification compared with standard classification tree building;
    ($\star$) at each node $d$, rather than choosing the best split among all variables, randomly select $m$ of the $p$ predictive variables;
    for $j = 1$ to $m$ do
        if $j$ is continuous then
            find the best cutpoint $t_j$ among all possible cutpoints for predictive variable $j$;
        else if $j$ is categorical then
            find the best categorical cutpoint $t_j$ among all classes for predictive variable $j$;
        end
    end
    select predictive variable $j$ with the lowest impurity at its best cutpoint $t_j$ to define the splitting rule of node $d$;
    if $j$ is continuous then
        send elements with $x_j < t_j$ to the left descendant, and elements with $x_j \geq t_j$ to the right descendant;
    else if $j$ is categorical then
        send elements with $x_j = t_j$ to the left descendant, and elements with $x_j \neq t_j$ to the right descendant;
    end
    repeat from ($\star$) on all descendant nodes to grow a classification tree to the maximal depth;
end
```

For random forest model development, Random Forests Version 5.1 [216] was used. The randomForest package within the statistical software R 2.2.1 [215] can also be used.

Algorithm 2: Computing the out-of-bag error (oob error).

```
Data: training data set $X$
Result: out-of-bag error of the random forest
define parameters $m$ and $k$;
for $i$ = 1 to $k$ do
    each classification tree is constructed using a different bootstrap sample $X_i$ from the original data set $X$. $X_i$ consists of about 2/3 of the elements of the original data set. The elements not included in $X_i$, called out-of-bag elements, are not used in the construction of the $i$–th tree;
    these out-of-bag elements are classified by the finalized $i$–th tree;
end
calculate the out-of-bag (oob) error as the proportion of misclassifications [%] over all out-of-bag elements.
```

4.3.4 Training versus test data sets

The lack of an independent data set for model evaluation forced to apply cross-validation (Algorithm 3). Here, in 2-fold cross-validation, each of two disjoint parts is once used as training set and once as test set:

Algorithm 3: Model construction and testing using q-fold cross-validation.

```
Data: data set $L$
Result: $q$-fold cross-validated model
define $q$;
randomly and uniformly split the ecohydrological data set $L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}$ into $q$ disjoint test data sets $L_{\text{test}i}$ $(i = 1, \ldots, q)$;
for $i$=1 to $q$ do
    use $L_{\text{train}i} = L - L_{\text{test}i}$ as training data set to construct the model;
    apply the model to test data set $L_{\text{test}i}$;
end
```

$$L_{\text{train1}} = L_{\text{test2}} \quad \text{of size } 853, \tag{4.17}$$

$$L_{\text{train2}} = L_{\text{test1}} \quad \text{of size } 852. \tag{4.18}$$

Consequently, each element $(\mathbf{x}_i, l_i)$ of the ecohydrological data set was once used as a training instance and once as a test instance.

4.4 Model construction, calibration and results

4.4.1 Multiple logistic regression model

The need to split the data set into two parts in order to cross-validate the results resulted in the construction of two multiple logistic regression models MLR1 and MLR2, constructed on L_{train1} and L_{train2} respectively. Additionally, the need to have a binomial response (0/1) for each vegetation type forced to redesign training and test data sets with multinomial (eleven vegetation types) response into eleven data sets with binomial response. Therefore, each of these models (MLR1 and MLR2) consisted of eleven submodels, i.e. estimated logit link functions $\hat{g}(\mathbf{x})$, one for each vegetation type. The submodels were constructed separately in two steps: (i) submodel construction using all 14 variables as first order terms and quadratic model terms, and (ii) bi-directional model term selection in a stepwise fashion using the AIC criterion. Casewise Pearson residual values (Eq. (4.15), [210]) were used to identify anomalous elements in the training set (elements with a Pearson residual > 15). These elements were excluded from the training set and the submodel building was repeated on the remaining training elements ($L'_{\mathrm{train1}} = 811$ training elements, and $L'_{\mathrm{train2}} = 812$ training elements). Algorithm 4 gives the algorithm for the construction of MLR1; MLR2 was constructed similarly using training data set L_{train2}. Indications on model goodness-of-fit are given in Table 4.6. Null model deviances (intercept only), residual deviances, likelihood ratio test G and Pearson chi-square are tabulated for MLR1 and MLR2. Since these statistical measures follow a χ^2–distribution, this distribution is used to test upon. The residual deviances were all smaller than null deviances, and therefore, the residual multiple logistic regression models were concluded to fit better than the null models. In order to statistically determine the degree of improvement, the likelihood ratio test statistic G was applied, which indicated that all multiple logistic regression models including significant predictive variables (as determined by the AIC criterion) fitted the observed vegetation type distribution better than the null models (intercept only) at the 0.01 significance level. The deviance goodness-of-fit (D_{resid}) and Pearson chi-square statistic (X_p^2) showed a significant fit between observations and fitted values at the 0.01 significance level[3].

After model construction, MLR1 was applied to L_{test1}, and MLR2 to L_{test2}. The joint output of MLR1 and MLR2 included the probability of occurrence $\hat{\pi}(\mathbf{x})$ for all eleven vegetation types for each measurement vector $\mathbf{x}$ in L and thus for each grid cell of the study area. The probabilities of occurrence $\hat{\pi}(\mathbf{x})$ for the eleven

---

[3] It is important, however, to note that observations are auto-correlated. This implies that the observations are not entirely independent [217], while the null deviance, residual deviance and Pearson chi-square statistic assume independence. The tabulated degrees of freedom (Df) for these statistics are therefore too high, and p-values and significance testing should therefore be interpreted with caution. The same comment should be made for Subsection 5.3.3.1, where the same statistics were applied, and for Section 5.3.1, where correlations between environmental variables are tested on significance.

Algorithm 4: Construction of the multiple logistic regression model MLR1.

```
Data: training data set $L_{\text{train1}}$
Result: calibrated model MLR1
for $i$ = 1 to 11 do
    make multinomial response of $L_{\text{train1}}$ binomial for each vegetation type;
    $L'_{\text{train1},i} = L_{\text{train1}}$ with binomial response;
end
for $i$ = 1 to 11 do
    construct submodel mlr$_i$ on training data set $L'_{\text{train1},i}$;
    select model terms stepwise using AIC;
    calculate $D_{\text{null},i}$, $D_{\text{resid},i}$ and $G_i$;
    count number of elements in $L'_{\text{train1},i}$ and assign to $n$;
    for $j$ = 1 to $n$ do
        calculate Pearson residual$_j$;
        if Pearson residual$_j$ > 15 then
            exclude $j$ from $L'_{\text{train1},i}$;
            go back to construct submodel mlr$_i$ on training data set $L'_{\text{train1},i}$;
        end
    end
end
group submodels mlr$_1, \ldots,$mlr$_{11}$ into MLR1;
```

different vegetation types do not necessarily sum up to 1 per grid cell, because the logit link functions $\hat{g}(\mathbf{x})$ were calculated separately for the eleven vegetation types. Based on a simple *decision rule*, i.e. *for each grid cell, the vegetation type with the highest probability of occurrence is the predicted vegetation type*, spatially distributed predictions of vegetation type occurrences were made (Fig. 4.7(a)). Out of the 1705 grid cells, 1182 (69.3%) were predicted correctly, 524 (30.7%) incorrectly. Visual inspection of the results (Fig. 4.7(a)) led to the conclusion that: (i) predictions were good for sites with little vegetation type diversity (Zwarte Beek); (ii) considerable numbers of predictions did not coincide with observations for the other, more diverse sites; and (iii) within the diverse sites, predictions were much better for large homogeneous vegetation clusters (e.g. northern area of Vorsdonkbos-Turfputten). However, for small and isolated patches and for boundary grid cells between neighbouring vegetation types, predictions were less accurate.

TABLE 4.6 – Model goodness-of-fit.

Vegetation type	D_{null}	Df	D_{resid}	Df	$G = D_{\text{null}} - D_{\text{resid}}$	Df	X_p^2	Df
MLR1								
Alno–Padion	472.01	810	107.15*	789	364.86*	21	130.68*	789
Arrhenatherion elatioris	548.49	810	173.10*	791	375.39*	19	403.18*	791
Calthion palustris	548.49	810	150.02*	793	398.47*	17	262.20*	793
Carici elongatae – Alnetum glutinosae	665.72	810	354.11*	799	311.61*	21	364.39*	799
Caricion curto–nigrae	581.79	810	0*	790	581.79*	20	0*	790
Cirsio – Molinietum	124.92	810	0*	798	124.92*	12	0*	798
Filipendulion	813.87	810	165.94*	794	647.93*	16	385.33*	794
Phragmitetalia	282.24	810	95.07*	803	187.17*	7	89.31*	803
Magnocaricion with *Phragmites*	539.91	810	133.25*	795	406.66*	15	176.49*	795
Magnocaricion	300.75	810	69.06*	795	231.69*	15	92.14*	795
Sphagno – Alnetum glutinosae	256.70	810	88.15*	800	168.55*	10	95.42*	800
MLR2								
Alno–Padion	513.73	811	134.01*	789	379.72*	22	122.98*	789
Arrhenatherion elatioris	452.92	811	184.44*	788	268.48*	23	235.94*	788
Calthion palustris	617.69	811	166.77*	796	450.92*	15	256.74*	796
Carici elongatae – Alnetum glutinosae	683.70	811	388.31*	795	295.39*	16	384.82*	795
Caricion curto–nigrae	609.93	811	13.62*	791	596.31*	20	15.18*	791
Cirsio – Molinietum	141.45	811	22.30*	790	119.15*	21	25.31*	790
Filipendulion	788.81	811	259.70*	791	529.11*	20	387.42*	791
Phragmitetalia	236.89	811	69.49*	795	167.4*	16	84.85*	795
Magnocaricion with *Phragmites*	222.39	811	133.25*	793	89.14*	18	254.34*	793
Magnocaricion	318.85	811	84.48*	795	234.37*	16	109.50*	795
Sphagno – Alnetum glutinosae	282.33	811	92.21*	789	190.12*	22	90.10*	789

D_{null} = deviance of the null model (intercept only model); Df = degrees of freedom; D_{resid} = residual deviance; G = the likelihood ratio test; X_p^2 = Pearson chi-square goodness-of-fit. Significance at the 0.01 level (*) are indicated for the residual deviance, likelihood ratio and the Pearson chi-square goodness-of-fit tests.

4.4.2 Random forest model

A schematic overview of random forest model construction and testing is given in Fig. 4.5. For a more detailed overview of the random forest model construction, the reader is referred to Algorithms 1 and 2. The random forest technique has two important user-defined parameters: the number of trees (k) and the number of randomly selected variables to split the nodes (m). These parameters should be optimized in order to minimize the generalization error, which is a machine learning function used to investigate the machine learning algorithm (here, the random forest algorithm) performance through iteration of the learning process (here, the calculation of additional classifiers (increase in k) and the increase of variebles to split the nodes (m)).

Breiman [204] proved that random forests do not overfit. A limiting value of the generalization error is obtained as more trees are added. Two random forest submodels RF1 and RF2 consisting of 10000 trees were constructed on L_{train1} and L_{train2} respectively, both with two randomly selected variables to split the nodes ($m = 2$). Fig. 4.6 presents the error in function of the number of trees. Two distinct forms of curves are distinguishable: (i) oob error and (ii) test set error. RF1 oob error and RF2 oob error represent the oob error, which was proven to be a good estimator of the generalization error [204], in function of the number of trees. From approximately 100 trees onwards, the oob error converged to about 20% for RF1, and to about 25% for RF2. Adding more trees did not decrease nor increase the oob error. The two other curves represent the test set error in function of the number of trees. Test set error values for different numbers of trees were computed by applying RF1 and RF2 to L_{test1} and L_{test2} respectively, during the random forest building process, and represent the proportion of incorrectly predicted test set elements. Test set error values for both test sets were around 23% at the end of the random forests construction. Similarly as for the oob error, the test set error converged from 100 trees onward. The conclusions that could be drawn from Fig. 4.6 are: (i) the oob error is a suitable estimator to detect error convergence, (ii) in accordance with Breiman [204] the random forest algorithm does not overfit: a limiting value for both oob error and test set error is produced, and (iii) 1000 trees can be concluded to be an appropriate size for both random forests in this study.

As stated in the random forest description, an additional random factor is included in the random forest algorithm compared with usual classification tree building: at each node a random subset of m predictive variables has to be specified and the best splitting variable among those m is used to split the node. The value of m is constant during the forest growing. It affects both the correlations between the trees and the strength of the individual trees. Reducing m reduces correlation and strength, increasing m increases both. Two random forests RF1 and RF2 were constructed for different values of m. Error values are tabulated in Table 4.7. Both the oob error for RF1 and RF2 constructed on L_{train1} and L_{train2} respectively, and

1. Training and test data sets

L_{train1}	
$\mathbf{x}_1$	🟥 l_1
$\mathbf{x}_2$	🟦 l_2
$\mathbf{x}_3$	🟦 l_3
$\mathbf{x}_4$	🟥 l_4
$\mathbf{x}_5$	🟨 l_5
$\mathbf{x}_6$	🟩 l_6
$\mathbf{x}_7$	🟨 l_7
$\mathbf{x}_8$	🟦 l_8
$\mathbf{x}_9$	🟥 l_9
$\vdots$	$\vdots$
$\mathbf{x}_V$	🟩 l_V

L_{test1}	
$\mathbf{x}_1$	🟨 l_1
$\mathbf{x}_2$	🟥 l_2
$\mathbf{x}_3$	🟦 l_3
$\mathbf{x}_4$	🟥 l_4
$\mathbf{x}_5$	🟩 l_5
$\mathbf{x}_6$	🟩 l_6
$\mathbf{x}_7$	🟥 l_7
$\mathbf{x}_8$	🟨 l_8
$\mathbf{x}_9$	🟥 l_9
$\vdots$	$\vdots$
$\mathbf{x}_W$	🟦 l_W

2. Construction of the random forest (RF1) consisting of *k* classification trees

(2a) take i ($i = 1, \ldots, k$) bootstrap subsamples from L_{train1}

	bootstrap sample	oob sample
1	bootstrap$_1$	oob$_1 = L_{\text{train1}} - \text{bootstrap}_1$
2	bootstrap$_2$	oob$_2 = L_{\text{train1}} - \text{bootstrap}_2$
$\vdots$	$\vdots$	$\vdots$
k	bootstrap$_k$	oob$_k = L_{\text{train1}} - \text{bootstrap}_k$

(2b) use the k bootstrap samples to construct k classification trees

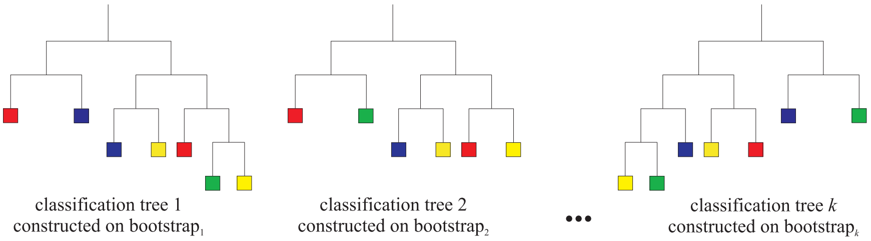
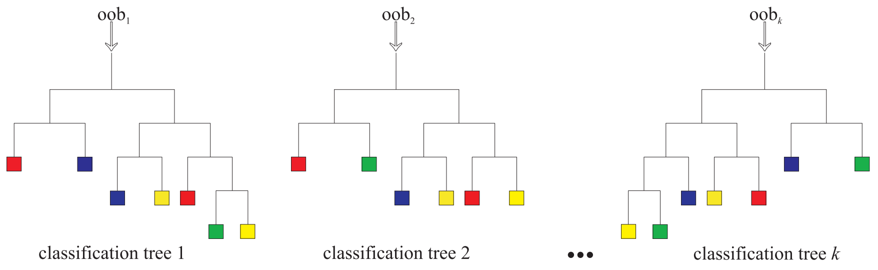

classification tree 1
constructed on bootstrap₁
classification tree 2
constructed on bootstrap₂
•••
classification tree k
constructed on bootstrapₖ

FIGURE 4.5 – Schematic overview of the construction of the random forest model, and its application on a test data set (example of RF1, the construction and testing setup of RF2 is identical).

(2c) apply classification tree$_i$ to oob$_i$ ($i = 1, \ldots, k$)



3. Apply the random forest (RF1) to the test data set L_{test1}

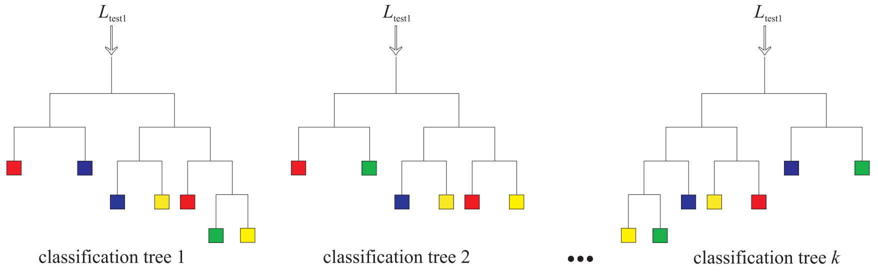


FIGURE 4.5 – *continued...*

test set errors for RF1 and RF2 applied to L_{test1} and L_{test2} respectively, are given.

The oob error showed minimal values of 19.91% for RF1 and 24.38% for RF2, both when $m = 3$. The test set error for RF1 applied to L_{test1} ranged between a minimum of 22.74% for $m = 5$ variables and a maximum of 25.32% for $m = 14$ variables. For RF$_2$ applied to L_{test2} similar error values were found for the different m values. A minimum of 23.42% was found for $m = 3$ and a maximum of 25.17% for $m = 14$. Overall, low oob error and test set error values were observed for $m = 3$. Therefore the oob error proved to be a good tool for optimizing m. In general little difference in error was found for $m \in \{2, 3, 4, 5, 8\}$. The optimal range of m was concluded to be quite wide (in accordance with Breiman and Cutler [218]). Nevertheless, it was decided to construct RF1 and RF2 with $m = 3$.

Based on the above findings (i.e. 1000 is a suitable number of trees and $m = 3$ results in a minimal error), the random forest algorithm was run on L_{train1} to create RF1 consisting of 1000 classification trees with three random predictive variables to split the nodes ($m = 3$). The same was done on L_{train2} to create RF2. Next, both random forests were applied to test data sets: RF1 on L_{test1} and RF2 on L_{test2}.

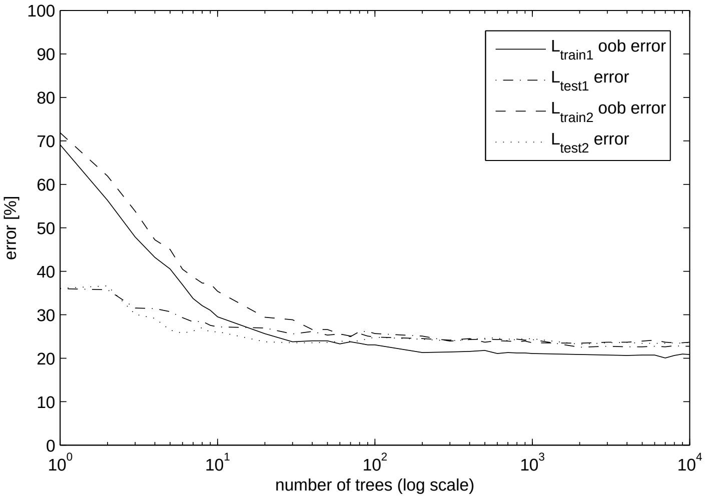


FIGURE 4.6 – Out-of-bag (oob) error and test set error converge when more trees are added to the random forest. L_{train1} oob error and L_{train2} oob error are the oob errors calculated during the construction of RF1 and RF2 respectively. L_{test1} error and L_{test2} error are the test set error of RF1 and RF2 applied to their respective test data sets.

TABLE 4.7 – Oob error values for RF1 and RF2 built on L_{train1} and L_{train2} respectively. Test set error values for RF1 and RF2 applied to L_{test1} and L_{test2} respectively.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=8$	$m=11$	$m=14$
^{a}RF1	21.78	20.26	$\underline{19.91}$	20.37	20.61	20.02	20.37	21.19
^{a}RF2	26.73	24.62	$\underline{24.38}$	24.85	24.62	24.38	24.62	$\underline{24.38}$
$^bL_{\text{test1}}$	24.62	23.33	23.33	23.33	$\underline{22.74}$	23.68	24.38	25.32
$^bL_{\text{test2}}$	25.06	23.77	$\underline{23.42}$	23.77	24.24	24.36	24.71	25.17

^{a} = oob error, ^{b} = test set error. Minimal values are underlined.

Each measurement vector $\mathbf{x}_i$ of the test sets was classified by each of the k trees in the ensemble as a unique vegetation type $c_j \in \{c_1, \ldots, c_{11}\}$. Consequently, each measurement vector $\mathbf{x}_i$ of the test sets is classified 1000 times and the proportion of votes over all 1000 trees for a vegetation type is interpreted as the probability of occurrence of that vegetation type:

$$P(c_j) = N_{c_j}/N_{\text{tot}}, \tag{4.19}$$

with $P(c_j)$ is the probability of occurrence of vegetation type c_j, N_{c_j} the number of trees classifying the vegetation type as vegetation type c$_j$, and N_{tot} $(= k)$ the total number of classification trees in the random forest (here $N_{\text{tot}} = 1000$).

This probability of occurrence was calculated for the eleven different vegetation types for each grid cell in the four study sites. The same *decision rule* as in multiple logistic regression modelling was used: *for each grid cell the vegetation type with the highest probability of occurrence is the predicted vegetation type*. Predictions were correct in the central area of all vegetation types (Fig. 4.7(b)). Predictions for grid cells at the boundary between different vegetation types and isolated cells were less accurate. Nonetheless, with 1307 (76.7%) correct predictions and 398 (23.3%) wrong predictions, the overall prediction accuracy was better than the prediction accuracy of the multiple logistic regression model which made 1182 (69.3%) correct predictions and 524 (30.7%) incorrect predictions.

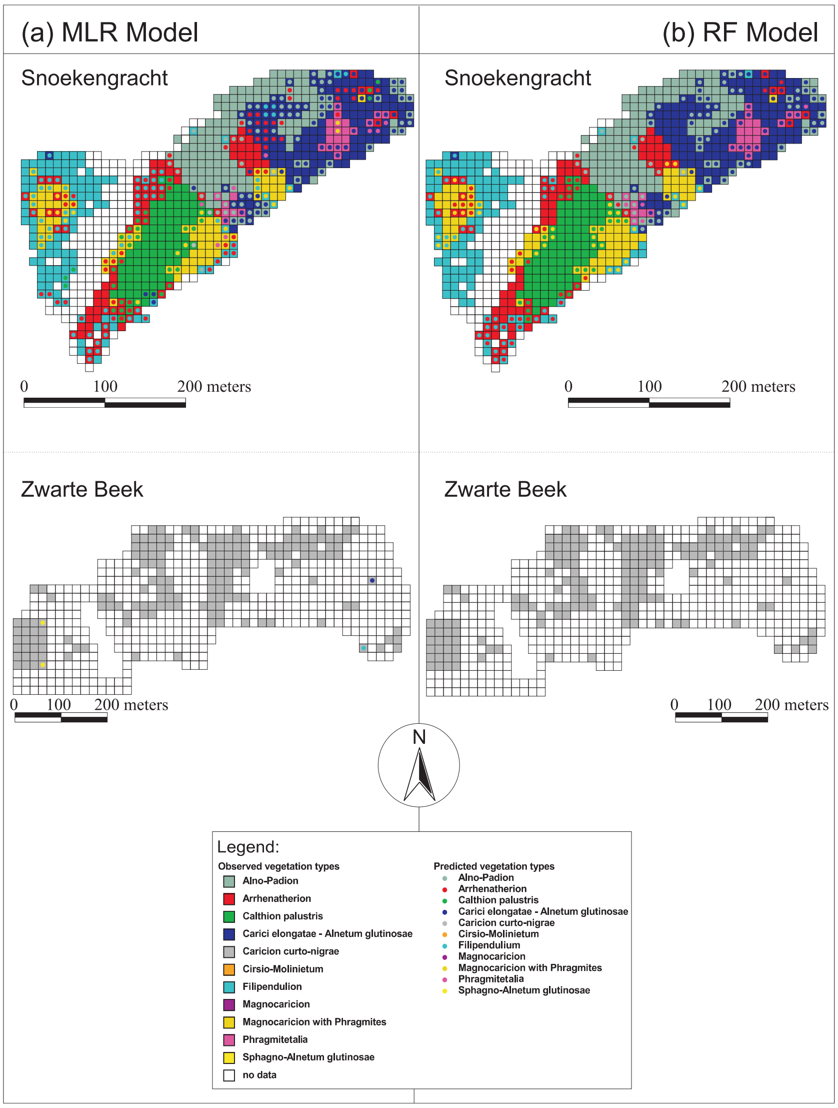


FIGURE 4.7 – Predicted vegetation types with the multiple logistic regression model (a) and with the random forest model (b). The observed vegetation distribution (☐) is overlaid with the predicted vegetation distribution (○). For each grid cell, the vegetation type with the highest probability of occurrence, as modelled with the multiple logistic regression model (a) and with the random forest model (b), is the predicted vegetation type.

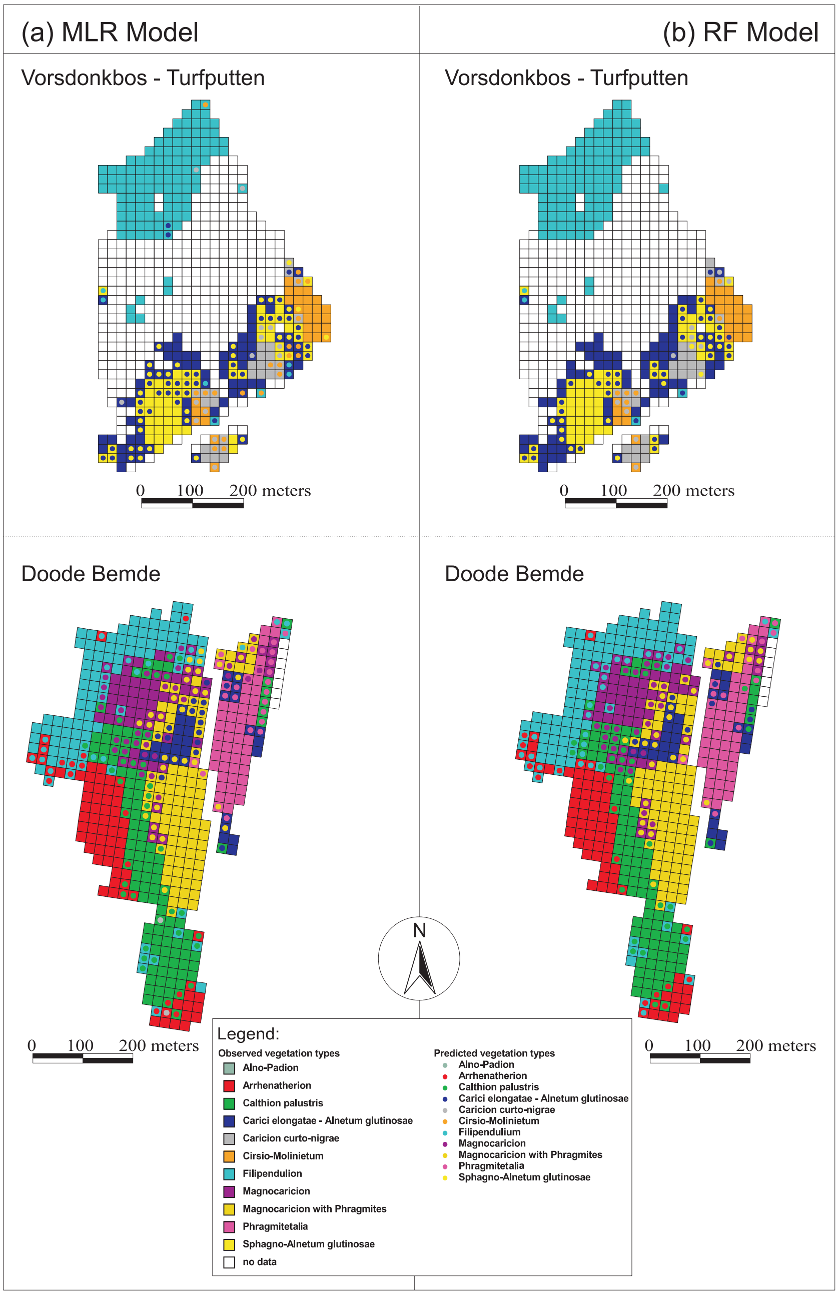


FIGURE 4.5 – *continued…*

4.5 Model evaluation

4.5.1 Observed versus predicted

The multiple logistic regression model and the random forest model consisted of two submodels: MLR1 and MLR2, and RF1 and RF2 respectively. This split resulted from 2–fold cross-validation. Vegetation type occurrences were predicted by applying MLR1 to L_{test1}, MLR2 to L_{test2} and RF1 to L_{test1}, RF2 to L_{test2}. From this point on, the joined predictions of the two parts of each model will be referred to as predictions made by the multiple logistic regression model and the predictions made by the random forest model. The performance of both models is discussed in this model evaluation section using different techniques.

Cohen's kappa test
Despite its weaknesses [219], the Cohen's κ test [220] was used to evaluate differences between observations and predictions. A confusion matrix was constructed in which observed and predicted vegetation types are given for each grid cell using the multiple logistic regression model (Table 4.8(a)) and the random forest model (Table 4.8(b)). For each of the confusion matrices (Table 4.8) the Cohen's kappa was calculated as [220]:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.20}$$

where P_o is the proportional observed agreement, and P_e the proportional agreement expected by chance. P_o and P_e are calculated as

$$P_o = \sum_{i=1}^{c} P_{ii} = \sum_{i=1}^{c} n_{ii}/N \quad \text{(here with } c = 11 \text{ vegetation classes)}$$

$$P_e = \sum_{i=1}^{c} P_{\cdot i} \cdot P_{i\cdot} = \sum_{i=1}^{c} [n_{\cdot i} \cdot n_{i\cdot}]/N \quad \text{(here with } c = 11 \text{ vegetation classes)}$$

where N is the total number of elements (here, 1705 grid cells), n_{ii} the number of elements in the diagonal cell ii, $n_{\cdot i}$ and $n_{i\cdot}$ are the totals of column i and row i, respectively. κ values are negative when the agreement between observations and predictions is worse than expected by chance, and reaches 1 in case of perfect agreement. A κ value of 0.651 was found for the multiple logistic regression model: there is a substantial agreement between observations and predictions ($p < 0.001$). A κ value of 0.734 was found for the random forest model: there is a substantial agreement between observations and predictions ($p < 0.001$). This κ value is higher than the one found for the multiple logistic regression model.

McNemar test
For $L = L_{\text{test1}} \cup L_{\text{test2}}$ (1705 elements spatially covering the whole study area) 1182 correct predictions were made by the multiple logistic regression model. The random forest model made 1307 correct predictions. Based on the conclusions

TABLE 4.8 – Model performances represented by confusion matrices in which observed vegetation types are compared with predicted vegetation types using the multiple logistic regression model (a) and the random forest model (b).

(a) Multiple logistic regression model

		observed										
		AP	*Ar*	*Cp*	*Ce*	*Cc*	*CM*	*Fi*	*Ma*	*MP*	*Ph*	*SA*
predicted	*AP*	111	7	0	33	0	0	3	0	0	4	0
	Ar	6	89	11	18	0	0	22	0	12	1	0
	Cp	0	19	156	2	0	0	13	7	14	0	0
	Ce	28	5	3	136	2	0	5	3	13	13	30
	Cc	0	1	3	2	181	5	2	0	0	0	5
	CM	0	0	0	5	11	21	2	0	0	0	2
	Fi	0	37	6	10	1	0	272	3	19	1	1
	Ma	0	0	10	3	0	0	6	29	8	2	0
	MP	1	3	6	6	0	0	9	16	105	5	0
	Ph	1	0	6	10	0	0	1	5	5	54	0
	SA	0	0	0	25	3	2	0	0	0	0	28

(b) Random forest model

		observed										
		AP	*Ar*	*Cp*	*Ce*	*Cc*	*CM*	*Fi*	*Ma*	*MP*	*Ph*	*SA*
predicted	*AP*	113	7	0	27	0	0	2	0	0	5	0
	Ar	3	102	11	2	0	0	18	0	11	0	0
	Cp	0	11	154	1	0	0	9	5	12	1	0
	Ce	28	5	3	189	1	1	3	0	7	15	29
	Cc	0	0	0	2	189	4	0	0	0	0	4
	CM	0	0	0	0	6	23	0	0	0	0	0
	Fi	0	29	4	2	0	0	286	9	5	1	1
	Ma	0	0	14	2	0	0	5	34	8	0	0
	MP	1	7	13	5	0	0	11	13	132	5	0
	Ph	2	0	2	6	0	0	1	2	1	53	0
	SA	0	0	0	14	2	0	0	0	0	0	32

of [221], the McNemar test [222] was selected to compare the performances of the multiple logistic regression model and the random forest model. Predictions made by both models for all cases of L (as presented in Fig. 4.7) were compared with the observations and used to construct the following contingency table (Table 4.9) where $N = n_{00} + n_{01} + n_{10} + n_{11}$ is the total number of elements in the ecohydrological data set (Table 4.9). Under the null hypothesis, the two models should have the same error rate, which means that $n_{01} = n_{10}$. McNemar's test is based on a χ^2–test for goodness-of-fit that compares the distribution of counts under the null hypothesis to the observed counts. The following statistic is χ^2–distributed with 1 degree of freedom:

$$M = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}. \tag{4.21}$$

TABLE 4.9 – Contingency table for the McNemar test.

number of grid cells misclassified by both MLR and RF n_{00}	number of grid cells misclassified by MLR but not by RF n_{01}
number of grid cells misclassified by RF but not by MLR n_{10}	number of grid cells misclassified neither by MLR nor by RF n_{11}

If the null hypothesis is correct, then the probability that this quantity is greater than $\chi^2_{0.95,1}$ = 3.84 is less than 0.05. Over the entire study area $n_{01} = 216$ and $n_{10} = 91$. The value of the test statistic M was 50.1 ($p < 0.001$). The two models had significantly different performances at the 0.001 significance level. Inspecting the n_{01} and n_{10} values led to the conclusion that this significant difference in performance was due to a better performance of the random forest model compared with the multiple logistic regression model.

Evaluation statistics for each vegetation type separately
To assess and compare model performances for each individual vegetation type, different test statistics were used. First, the McNemar test was used to identify differences in performance of both models for each vegetation type separately. Furthermore, predicted vegetation types by the two models were compared with observed vegetation types for the eleven vegetation types separately using a confusion matrix (see also further in Table 6.2, e.g. [223, 224]):

TABLE 4.10 – Confusion matrix. TP stands for True Positive, FP for False Positive, FN for False Negative and TN for True negative.

		observed	
		present	absent
predicted	present	TP	FP
	absent	FN	TN

In such a confusion matrix, observations are compared with model predictions for each vegetation type separately. Given a test instance (grid cell from cross-validation test data set), there are four possibilities with respect to the vegetation type of interest: (i) if the observed and modelled vegetation type of a test instance coincide, and this vegetation type is the one of interest, the test instance is counted for as true positive (TP), (ii) if the observed and predicted vegetation type of a test instance do not coincide, and the predicted vegetation type is the one of interest, the test instance is counted for as false positive (FP), (iii) if the observed and predicted vegetation type of a test instance do not coincide, and the observed vegetation type is the one of interest, the test instance is counted for as false negative (FN), and (iv) if the observed and predicted vegetation type of a test instance coincide, but this vegetation type is not the one of interest, the test instance is counted for as true negative (TN).

Using these four possible outcomes (TP, FP, FN, TN), several standard terms have been defined for a confusion matrix [223, 224] of which following were used because of our main interest in correctly predicting presences:

(i) *Precision*, p (=positive predictive power): the proportion of predicted presences that are observed to be present rather than absent, TP/(TP + FP);

(ii) *Recall*, r (=sensitivity, =true positive rate): the proportion of observed presences that were predicted correctly, TP/(TP + FN)

Precision and recall were combined by means of the ‘F-measure’ [225]. A weighted version of the F-measure was used:

$$F_\beta(\mathrm{p},\mathrm{r}) = \frac{(\beta^2+1)\mathrm{pr}}{\beta^2\mathrm{p}+\mathrm{r}}, \tag{4.22}$$

where $\beta \in]0,+\infty[$ is a weighing factor that controls the relative importance of precision versus recall. For $\beta = 1$, the F-measure is balanced, and precision and recall have equal importance. The F-measures used were $F_{0.5}$ (precision twice as important as recall), F_1 (equal weights) and F_2 (recall twice as important as precision). The magnitude of F varies from 0, when all observed presences are predicted incorrectly, to 1, when predictions and observations perfectly match. Moreover F is strongly oriented towards the lower of the two values p and r; therefore this measure can only be high when both p and r are high.

Results of the McNemar test and values for precision, recall and the F-measure are summarized in Table 4.11 for the individual vegetation types. The F-measures for the two models over all vegetation types were analysed using two test statistics: (i) a simple ranking and (ii) the Wilcoxon signed rank test. Simple ranking assigned performance scores per vegetation type: 2 for the best performing model and 1 for the worst and 1.5 in case of a tie. After adding up those values for each of the F-measures, the highest scoring model was concluded to perform best. The Wilcoxon signed rank test [226] is a non-parametric pairwise comparison test. It allows to test whether the median values of the different F-measures over the different vegetation types are identical for the two models.

The McNemar test showed a significant difference in performance between the multiple logistic regression model and the random forest model at the 0.05 significance level for the vegetation types *Arrhenatherion elatioris*, *Carici elongetae – Alnetum glutinosae*, *Caricion curto–nigrae*, *Filipendulion* and *Magnocaricion*

TABLE 4.11 – McNemar test for comparison of MLR and RF model performances for each vegetation type individually. Precision, recall, and three F-measures ($F_{0.5}$, F_1, F_2) for MLR and RF modelling results are given for each vegetation type separately as well.

		Vegetation type										
		Alno–Padion	*Arrhenatherion elatioris*	*Calthion palustris*	*Carici elongetae – Alnetum glutinosae*	*Caricion curto–nigrae*	*Cirsio–Molinietum*	*Filipendulion*	*Magnocaricion*	*Magnocaricion with Phragmites*	*Phragmitetea*	*Sphagno–Alnetum glutinosae*
	McNemar test	n	y	n	y	y	n	y	n	y	n	n
	n_{01}	9	22	12	69	10	3	28	11	38	5	9
	n_{10}	7	9	14	16	2	1	14	6	11	6	5
MLR	precision	0.70	0.56	0.74	0.57	0.91	0.51	0.78	0.50	0.69	0.66	0.48
	recall	0.76	0.55	0.78	0.54	0.91	0.75	0.81	0.46	0.60	0.68	0.42
	$F_{0.5}$	0.71	0.56	0.75	0.57	0.91	0.55	0.78	0.49	0.67	0.66	0.47
	F_1	0.73	0.56	0.76	0.56	0.91	0.61	0.79	0.48	0.64	0.67	0.45
	F_2	0.74	0.55	0.77	0.55	0.91	0.69	0.80	0.47	0.61	0.67	0.43
RF	precision	0.73	0.69	0.80	0.67	0.95	0.79	0.85	0.54	0.70	0.79	0.67
	recall	0.77	0.63	0.77	0.75	0.95	0.82	0.85	0.54	0.75	0.66	0.48
	$F_{0.5}$	0.74	0.68	0.79	0.69	0.95	0.80	0.85	0.54	0.71	0.76	0.62
	F_1	0.75	0.66	0.78	0.71	0.95	0.81	0.85	0.54	0.73	0.72	0.56
	F_2	0.76	0.64	0.77	0.74	0.95	0.82	0.85	0.54	0.74	0.68	0.51

McNemar test: y = significant difference in performance between the MLR model and the RF model, n = no significant difference, both at the 0.05 significance level. n_{01} and n_{10} are error rates of the MLR model and the RF model respectively to calculate the McNemar test statistic M, see Eq. (4.21).

with *Phragmites*. These differences resulted from a better performance of the random forest model as can be seen from the n_{01} and n_{10} values in Table 4.11. The absence of significant differences between both models for the remaining vegetation types reflects comparable performances for both models due to a spatial distribution in large homogeneous areas for which predictions by both models are good (e.g. *Calthion palustris*, *Phragmitetea*) or due to spatial limitations of the vegetation type (e.g. *Alno–Padion* and *Magnocaricion* are only found at Snoekengracht and Doode Bemde respectively).

For precision and recall the same tendencies were noticeable for the two models. Precision for *Sphagno–Alnetum glutinosae* and *Magnocaricion* were low for both models, meaning that many cells with other vegetation types — mainly *Carici elongetae – Alnetum glutinosae* — were predicted to be *Spagno–Alnetum glutinosae* and many cells — mainly *Magnocaricion* with *Phragmites* and *Calthion palustris* — were predicted to be *Magnocaricion* (Fig. 4.7 and Table 4.8). This is somewhat understandable as these are spatially adjacent, comparable vegetation types with dominance of *Alnus glutinosa* (L.) Gaertn. in both *Sphagno–Alnetum glutinosae* and *Carici elongetae – Alnetum glutinosae*, and the higher abundance of *Phragmites australis* as main difference between *Magnocaricion* and *Magnocaricion* with *Phragmites* (see Section 3.2.2 and Table 3.2). Recall was lowest for *Sphagno–Alnetum glutinosae* and *Magnocaricion* for the multiple logistic regression and the random forest model. In Fig. 4.7 the large number of wrong predictions for *Sphagno–Alnetum glutinosae* and *Magnocaricion* in Vorsdonkbos-Turfputten and Doode Beemde are clearly noticeable. A similar explanation as for precision might be given. Many grid cells with observed *Sphagno–Alnetum glutinosae* and *Magnocaricion* vegetation were predicted to be the related vegetation type *Carici elongetae – Alnetum glutinosae* and *Magnocaricion* with *Phragmites*, respectively. Both models had high precision and recall for *Caricion curto–nigrae* probably resulting from well-defined differences of the environmental conditions, as concentrations of Mg^{2+}, Ca^{2+} and Cl^{-} are markedly lower at Zwarte Beek where this vegetation type was predominantly found (see Section 3.3.2.2).

The stated findings for precision and recall were reflected in the F-measures. F_1-values ranged between 0.45 and 0.91 for the multiple logistic regression model and between 0.56 and 0.95 for the random forest model. One-by-one comparison showed a better performance of the random forest model for all three F-measures for each of the eleven vegetation types. Based on the simple ranking statistic, all three F-measures were found to be better for the random forest model (11 for the multiple logistic regression model versus 22 for the random forest model). The Wilcoxon signed rank test statistic indicated significantly better performances for all three F-measures for the random forest model compared to the multiple logistic regression model at the 0.01 significance level ($p = 0.003$).

4.5.2 Prediction probabilities

Threshold dependent evaluation

The multiple logistic regression model and the random forest model computed the probabilities of occurrence for each individual vegetation type for each spatially distributed grid cell. Probability distributions for correct predictions and incorrect predictions gave an indication of the strength of the predictions (Fig. 4.8). Correct predictions were made with high probability, especially for the MLR model: half of the correct MLR model predictions had probabilities higher than 0.9, while one-third of the correct RF model predictions had probabilities higher than 0.9. A visual inspection of the probabilities underlying each prediction (not shown) indicated that correct predictions with high probabilities were found in the central areas of homogeneous vegetation clusters. Probabilities decreased toward the margins of those areas. Incorrect prediction probabilities tended to be rather high for the MLR model, with almost 20% of the incorrect predictions having higher probabilities than 0.9. Incorrect RF model prediction probabilities showed a maximum in the]0.4,0.5] interval indicating that incorrect predictions are mainly made for grid cells with several vegetation types with comparable, low to moderate probabilities. Only 2% of the incorrect predictions had probabilities higher than 0.9.

Spatial identification of these grid cells indicated them as isolated vegetation types, surrounded by other vegetation types.

Threshold independent evaluation
Receiver operating characteristic (ROC) curves are frequently used for the evaluation of classification accuracy [210, 227]. This curve, originating from signal detection theory, is widely used in clinical sciences, but recently also in earth sciences [51, 228–231]. ROC graphs are two-dimensional graphs in which the true positive rate (=recall), tpr, is plotted on the Y-axis, and the false positive rate, fpr, on the X-axis, with

$$\text{tpr} = \frac{\text{observed positives correctly classified}}{\text{total observed positives}} = \frac{\text{TP}}{\text{TP + FN}} \tag{4.23}$$

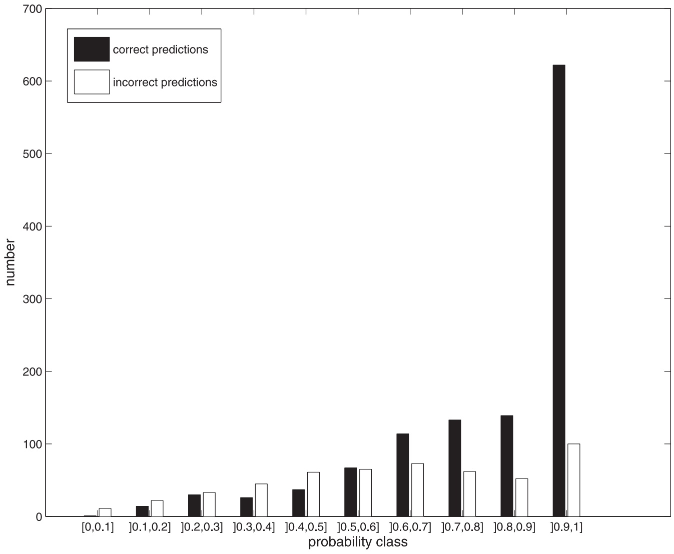
$$\text{fpr} = \frac{\text{observed negatives incorrectly classified}}{\text{total observed negatives}} = \frac{\text{FP}}{\text{FP + TN}}\,. \tag{4.24}$$

The true positive rate measures the fraction of observed presences (vegetation type present in grid cells) that are predicted correctly. The false positive rate measures the fraction of observed absences (vegetation type absent in grid cells) that are incorrectly predicted as present.

The multiple logistic regression model and the random forest model computed the probabilities of occurrence of eleven vegetation types. Earlier we used the decision rule that the most probable vegetation type (among the eleven possible vegetation types) is the predicted one. Here, in order to construct ROC curves for each vegetation type separately, the modelled probabilities of occurrence are used to construct several confusion matrices, one for each possible cutpoint. A cutpoint represents a threshold probability above which the vegetation type is modelled to be present. The curve generated by plotting the tpr versus the fpr for all possible cutpoints is the ROC curve. A simple example of how a ROC curve is generated is given in Fig. 4.9.

The area under the ROC curve (AUC), which ranges from zero to one, provides a measure of the ability of the model to discriminate between grid cells where the vegetation type of interest is present versus absent [210]. AUC describes the likelihood that the observed vegetation type for a grid cell has a higher modelled probability of occurrence in comparison with grid cells where the vegetation type is absent, and when the AUC value is higher than 0.5 the model does better than random guessing. Both models had high AUC–values, reflecting their excellent discrimination abilities (Table 4.12). *Alno – Padion* for example, has an AUC–value of 0.967 under the multiple logistic regression model, strongly indicating that grid cells in the study area where the *Alno – Padion* vegetation is present have a higher modelled probability of *Alno – Padion* occurrence than grid cells where *Alno – Padion* is absent. Nevertheless, the comparison of AUC–values over all vegetation types using Wilcoxon signed rank statistic indicated a significantly

(a) Multiple logistic regression model



(b) Random forest model

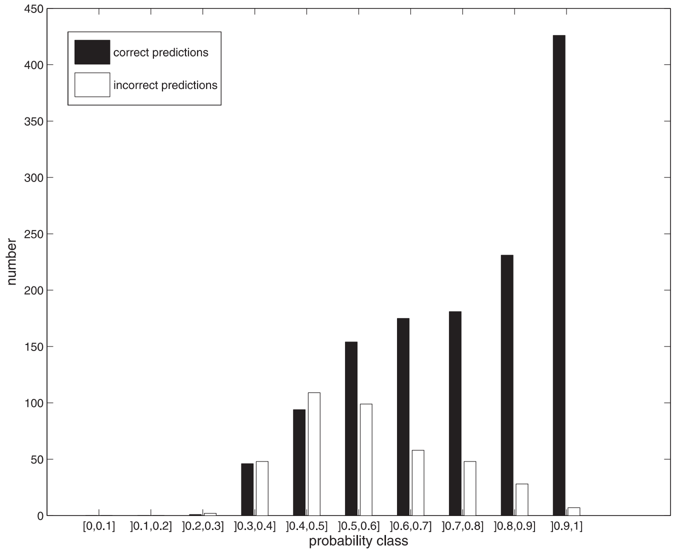


FIGURE 4.8 – Probability distributions of predictions made with the multiple logistic regression model (a) and the random forest model (b) ($N = 1705$).

nr.	observed	modelled probability $(P(c_j))$	A	B	C
1	present	0.8	present	present	absent
2	absent	0.7	present	present	absent
3	absent	0.1	present	absent	absent
4	present	0.6	present	present	absent
5	absent	0.3	present	absent	absent
6	present	0.9	present	present	absent
7	present	0.9	present	present	absent
8	present	0.4	present	absent	absent
9	absent	0.1	present	absent	absent
10	present	0.5	present	absent	absent

A

$P(c_j) = 0$

6	4
0	0

tpr=6/6=1
fpr=4/4=1

B

$P(c_j) = 0.5$

4	1
2	3

tpr=4/6=2/3
fpr=1/4

C

$P(c_j) = 1$

0	0
6	4

tpr=0/6=0
fpr=0/4=0

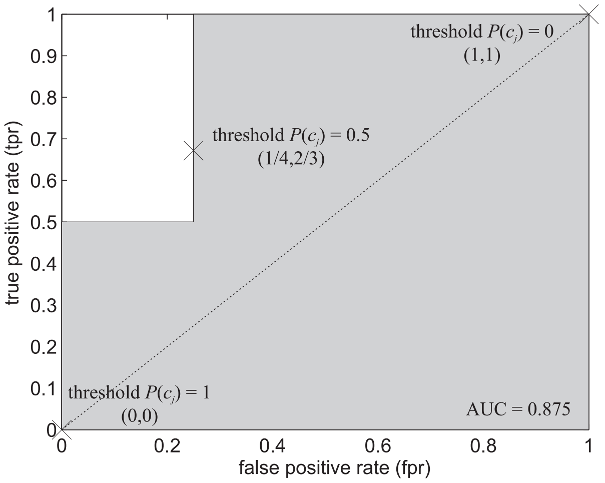


FIGURE 4.9 – Example of how a ROC graph is created. Ten test instances of which the presence/absence of a vegetation type is observed and the probability of occurrence is modelled $(P(c_j))$, are used to calculate the true positive rate (tpr) and false positive rate (fpr) of confusion matrices constructed by applying three different threshold probabilities: (A) $P(c_j) = 0$, (B) $P(c_j) = 0.5$, and (C) $P(c_j) = 1$. The ROC curve generated when all possible threshold probabilities are used is presented, and the three (fpr,tpr) pairs calculated are indicated (×). The dashed diagonal line (fpr = tpr) represents the ROC curve when vegetation types are classified by random guessing. The shaded area under ROC curve (AUC) has an area of 0.875, the AUC under the dashed ROC curve equals 0.5.

higher median AUC–values for the random forest model at the 0.01 significance level.

TABLE 4.12 – Area under ROC curves for the MLR and the RF model.

Vegetation type	MLR model	RF model
Alno – Padion	0.967*	0.983*
Arrhenatherion elatioris	0.920*	0.950*
Calthion palustris	0.927*	0.981*
Carici elongatae – Alnetum glutinosae	0.880*	0.949*
Caricion curto–nigrae	0.969*	0.999*
Cirsio – Molinietum	0.758*	0.886*
Filipendulion	0.923*	0.977*
Phragmitetalia	0.904*	0.963*
Magnocaricion with *Phragmites*	0.910*	0.969*
Magnocaricion	0.968*	0.983*
Sphagno – Alnetum glutinosae	0.950*	0.982*

* using the model for predicting vegetation type occurrence is better than random guessing at the 0.001 significance level.

4.6 Discussion and conclusions

4.6.1 Statistical model comparison

This study presented an application of two different predictive ecohydrological distribution models. The first model used the widely applied multiple logistic regression technique, and the second model a recently developed ensemble learning technique called random forest. Both models calculated the probability of occurrence of eleven different vegetation types, on which the prediction of the spatial vegetation distribution was based. An ecohydrological data set with hydrogeochemical variables and related vegetation types for Flemish lowland valley ecosystems was randomly and uniformly split into two training data sets for 2–fold cross-validation of both models. After model construction and calibration, the prediction accuracy of both models was assessed and compared. Following conclusions could be drawn:

1. The multiple logistic regression model made 69.3% correct predictions and the random forest model 76.7%. The McNemar test statistic indicated a difference in performance between the models at the 0.001 significance level ($p < 0.001$). Inspection of the results assigned this difference to a better performance of the random forest model compared to the multiple regression model.
2. The overall better performance of the random forest model could be assigned to significantly higher proportion of correct predictions for *Arrhenatherion elatioris*, *Carici elongetae – Alnetum glutinosae*, *Caricion curto–nigrae*, *Filipendulion* and *Magnocaricion* with *Phragmites* (see Table 4.11).

3. The F-measures, which combines precision and recall, were significantly better for the random forest model ($p = 0.003$).

4. The multiple logistic regression model made correct predictions with higher probabilities than the random forest model (Fig. 4.8). Unfortunately, the incorrect predictions were also made with high probabilities. The random forest model made incorrect predictions with lower probabilities, which indicated that the model misclassified grid cells where several vegetation types were expected, all with comparable, moderately low probabilities. Both models predicted central areas of homogeneous areas correctly with high probabilities, and isolated grid cells incorrectly with high probabilities.

5. Model accuracy was assessed by means of ROC curves for the vegetation types separately. The area under the curves (AUC) was high for both models, they were both much better for predicting vegetation occurrence than random guessing ($p < 0.001$). Although both models performed well, the random forest model was found to have higher discriminative power than the multiple logistic regression model at the 0.01 significance level.

The overall conclusion of this chapter is that the random forest modelling technique has the ability to lead to better predictive ecohydrological distribution models.

4.6.2 Putting the random forest model in a broader perspective

Major applications of the random forest classifier are found in bio-informatics and genetics (e.g. [232, 233]) and within the earth sciences in remote sensing (e.g. [234–236]). At the time this study was conducted, no example of the use of the random forest technique in ecological distribution modelling was found, and therefore comparison possibilities with literature were few. However, in two recent publications, Garzon et al. (2006, [190]) developed a random forest model to predict habitat suitability for Scots pine on the Iberian Peninsula, and Prasad et al. (2006, [189]) used the random forest technique to model future distributions of Loblolly pine, Sugar maple, American beech and White oak in North America under a climate change scenario. Both studies found superior distribution modelling performances of the random forest model compared with other techniques. Therefore the conclusion of this chapter can be generalized: the random forest modelling technique has the ability to lead to better distribution models for a variety of species and vegetation types in a variety of environments.

Nevertheless, general remarks on the random forest model should put its implementation within a broader perspective. As the random forest models statistically relate the occurrence of vegetation types to their present environment, the incorporation of functional relationships between environmental gradients and vegetation type distribution is not straightforward, and only partly possible in these empirical modelling approaches (this holds for the multiple logistic regression model as well). A first tendency towards more mechanistic modelling can be achieved by selecting causal variables (with direct physiological impact) as environmental variables. Austin [97] distinguished different classes of environmental gradients: (i) indirect gradients with no physiological effect on plant growth or competition (e.g. latitude or longitude); (ii) direct gradients with a direct physiological influence on growth without being consumed by plants (e.g. temperature and pH); and (iii) resource gradients including light, water and nutrients. The position of an environmental gradient in the chain of processes that link the gradient to its impact on the plant is either proximal or distal [97]. The most proximal gradient will be the causal variable determining the plant response. When proximal resources and direct gradients are used as environmental variables in modelling, the model will gain robustness and extend its range of applicability.

However, even if only proximal gradients were used in this modelling exercise, predictions would not completely fit the observations since ecological processes such as competition, predation and dispersal and other spatially autocorrelated features were not included. These processes tend to be hard to introduce into predictive models [237] because the actual vegetation type distribution is a result of both environmental conditions and ecological processes and their relative importance is hard to capture. Consequently, predictions made by the presented models are rather to be interpreted as habitat suitability maps for the different vegetation types [135].

In order to gain functionality of the random forest model, further research should focus on its modelling ability with smaller data subsets (see Chapter 5, comprising (most likely) uncorrelated proximal predictive variables. There are several reasons to do so [238]: (i) the model will gain robustness, with higher confidence on future predictions, (ii) some causal relationships can possibly be indicated and (iii) the utilization of the model would become less costly. Furthermore, model generality should be tested on a spatially independent data set since the use of accuracy estimates based on 2-fold cross-validation data and on spatially independent evaluation data tend to differ [151] (see Chapter 6).

5

Identification of important environmental variables in eco-hydrological distribution modelling

5.1 Introduction

Exploring the distribution of plant species and vegetation types is a central goal in ecology. Numerous studies have examined environmental gradients in relation to plant species or vegetation type distributions in various ecosystems (e.g. [81, 239–241]). Most modelling approaches developed for assessing species or vegetation type distributions have their roots in quantifying species-environment or vegetation-environment relationships [237]. Distribution models are mostly empirical models relating field observations to environmental variables based on statistically or theoretically derived responses [51]. Austin [97] distinguished different classes of environmental variables: (i) indirect variables with no physiological effect on plant growth or competition (e.g. latitude or longitude); (ii) direct variables with a direct physiological influence on growth without being consumed by the plant; and (iii) resource variables including light, water and nutrients. The position of an environmental variable in the chain of processes that link the variable to its impact on plants is either proximal or distal [97]. The most proximal

---

The content of this chapter is submitted as J. Peters, N. E. C. Verhoest, B. De Baets, R. Samson and P. Boeckx. *Wetland vegetation distribution modelling for the identification of constraining environmental variables* Landscape Ecology, accepted, 2008 (DOI : 10.1007/s10980-008-9261-4).

variables will causally determine plant responses.

A common feature of many species distribution models is that there are often many candidate predictive variables [99]. Additionally, variables are frequently significantly intercorrelated (multicollinearity) so that identifying the causal variables is problematic [242]. This large number of variables may result in overfitting with resulting models performing well in the context of the data set used to create them but not robust when applied elsewhere [99]. However, selecting the most influential variables in the model is not an easy task.

Multiple logistic regression within the framework of generalized linear models (GLM, [93]) is very popular and often used for modelling vegetation type distributions (Chapter 4 and e.g. [51, 100]). Within these modelling strategies stepwise selection procedures have been used for selecting the most influential variables. However, serious shortcomings have been reported [243] and new approaches have been proposed, such as hierarchical partitioning [242,244,245]. Another technique which has been applied in distribution modelling is random forests [204] (Chapter 4 and [131]). Within the random forest technique, the ‘variable importance’ measure is incorporated to determine the most influential variables.

As can be seen, a dichotomy in distribution modelling approach is prevalent: *predictive modelling* versus *explanatory modelling.* Predictive modelling aims at the development of distribution models while the focus is on the goodness-of-fit of the models. These models are typically ‘black-box’ models. The relative importance of the variables and the nature of their relationship with the vegetation distribution is of minor importance. Contrarily, explanatory modelling encompasses the exploration of measured environmental variables, with the intention to identify the most influential ones in explaining the vegetation distribution.

Based on these two approaches, a reduced distribution model using only the most important environmental variables can be constructed and evaluated on performance. This link between predictive and explanatory modelling is a priority for applied ecologists. The most influential variables, defined under explanatory modelling, can lead toward readily understandable distribution models while maintaining quantitative rigour with minimal resources [219].

This chapter formulates an answer to the research question:
Can the random forest distribution model provide information concerning environmental variable importance?

Additionally, two sub-questions are addressed:

a. Would other techniques identify the same variables as being important?
b. Is it possible to construct accurate random forest distribution models on a reduced data set, only including the most important environmental variables?

This chapter has a dichotomous structure, including (i) predictive modelling, and (ii) explanatory modelling. Predictive modelling is applied using multiple logistic regression and random forest distribution models. Explanatory modelling comprises the identification of important environmental variables, and is applied using three techniques: ordination, hierarchical partitioning of the multiple logistic regression models and the ‘variable importance’ measure within the random forest algorithm. Results from explanatory modelling are used to construct reduced models, and modelling results are compared with results from predictive modelling.

5.2 Material and methods

5.2.1 Test site and data set

From this chapter onwards, focus is exclusively on the Doode Bemde test site. A description of the area, together with the monitoring scheme of the site is given in Chapter 3. Groundwater depth measurements were used to calculate four groundwater quantity variables: average groundwater depth (AGD, [m]), maximal (Max) groundwater depth [m], minimal (Min) groundwater depth [m], and the amplitude (Ampli) of the groundwater depth [m]. Values of these variables, together with the groundwater quality variables, were assigned to each grid cell by spatial interpolation of measurement data over the entire area using block kriging (for details, see [78]). Mean values and standard deviations of the continuous environmental variables are summarized in Table 5.1.

The spatially explicit variables were structured into a data set. The data set contains $N = 519$ measurement vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ consisting of the values of $p = 17$ predictive variables describing the abiotic environment:

- Groundwater quantity: average groundwater depth, maximal groundwater depth, minimal groundwater depth, and the amplitude of the groundwater depth. All these variables are continuous;

TABLE 5.1 – Summary of the quantitative environmental variables as measured at the Doode Bemde. All values in [mg L^{-1}] except for average groundwater depth (AGD) [m], maximal groundwater depth (Max) [m], minimal groundwater depth (Min) [m], amplitude (Ampli) [m], pH [-] and soil organic matter content (SOM) [%]. The categorical variables soil type (loam, peat) and management (yearly, transition from yearly to cyclic (Y/C), cyclic, no management (no man.)) have no mean and no standard deviation [/]. Kendall correlation coefficients (τ) of the first two DCA and CCA axes are included. Four environmental variables were excluded from the CCA due to multicollinearity problems [–].

	Mean	Standard Deviation	τ with DCA axes		τ with CCA axes	
			Axis 1	Axis 2	Axis 1	Axis 2
AGD	-0.45	0.49	-0.429	-0.204	-0.449	-0.265
Max	-1.07	0.60	-0.541	-0.234	–	–
Min	-0.11	0.38	-0.260	-0.096	–	–
Ampli	0.96	0.36	0.506	0.291	0.547	0.157
pH	6.76	0.14	0.024	0.328	0.032	0.186
Cl^-	21.21	5.18	-0.378	0.105	-0.376	0.110
Ca^{2+}	99.99	24.58	-0.091	0.243	-0.087	0.346
Fe_{tot}	20.28	13.37	-0.181	-0.256	-0.198	-0.075
K^+	1.22	0.97	-0.195	-0.192	-0.236	-0.049
Mg^{2+}	6.75	1.25	0.094	0.233	0.101	0.141
NO_3^-–N	0.69	0.74	0.082	0.133	0.102	-0.022
NH_4^+–N	0.77	1.20	0.326	0.090	0.336	0.130
$H_2PO_4^-$	0.27	0.13	-0.060	-0.174	-0.078	-0.322
SO_4^{2-}	22.33	8.69	-0.414	-0.134	-0.434	-0.023
SOM	21.19	15.05	-0.587	-0.201	-0.599	-0.081
Loam	/	/	0.426	0.301	0.444	0.298
Peat	/	/	-0.426	-0.301	–	–
Yearly	/	/	0.540	-0.201	0.516	-0.156
Y/C	/	/	0.054	-0.023	0.053	-0.109
Cyclic	/	/	-0.337	0.299	-0.282	0.147
No man.	/	/	0.114	-0.281	–	–

- Groundwater quality: pH, Cl^-, Ca^{2+}, Fe_{tot}, K^+, Mg^{2+}, NO_3^-–N, NH_4^+–N, $H_2PO_4^-$ and SO_4^{2-}. All these variables are continuous;

- Soil: soil type (loam/peat, categorical), and soil organic matter content (continuous);

- Management: yearly mowing, cyclic mowing, transition from yearly to cyclic mowing, no management (categorical).

Seven different vegetation types $c_1, \ldots, c_7$ are considered: *Arrhenatherion elatioris*, *Calthion palustris*, *Carici elongetae – Alnetum glutinosae*, *Filipendulion*, *Phragmitetalia*, *Magnocaricion* with *Phragmites* and *Magnocaricion*, of which a short description is given in Chapter 3. To each measurement vector $\mathbf{x}_i$ a unique vegetation type $l_i \in \{c_1, \ldots, c_7\}$ is assigned. The data set will be denoted as:

$$L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\} \; . \tag{5.1}$$

5.2.2 Detrended correspondence analysis (DCA) and canonical correspondence analysis (CCA)

DCA [246] and CCA [193] were used for studying environmental gradients in relation to vegetation distributions at the Doode Bemde. DCA first ordinates species or vegetation data in an ordination diagram, which is then interpreted in the light of explicit environmental data [192]. This two-step approach is an indirect gradient analysis in the sense of Whittaker [247]. By contrast, CCA relates variation in species or vegetation to environmental variation directly, enabling the significant relationships between environmental variables and species or vegetation type distributions to be determined. The results obtained from both ordination techniques were compared as recommended by ter Braak [193] using the coefficient of determination (r^2) which is proportion of variability accounted for by a statistical model (here, ordination technique). r^2 takes values between 0 and 1, and equals 0 when no variability is accounted for, and 1 when all variability is accounted for. A larger r^2 value appears when more variability is accounted for. DCA and CCA were performed using PC–ORD Version 4 software, with downweighing of rare species and rescaling of the axes as selected options for the DCA.

5.2.3 Multiple logistic regression

Recalling the multiple logistic regression model description (see Eqs. (4.4) and (4.5)), an estimator $\hat{g}(\mathbf{x})$ for the logit function has to be found for each vegetation type. However, there are two difficulties [238]: (i) multiple regression is plagued by collinear relationships among predictive variables and (ii) any regression is designed to produce a function that in some way minimizes the overall difference between the observed and ‘predicted’ response values (here vegetation types), which does not necessarily imply causal dependence [248]. As Mac Nally [238] stated, the quest is to select q independent predictive variables from a set of p predictive variables (subset selection), subject to the problem of multicollinearity, because: (i) it is always possible to produce a better fit to the data by using more terms, and when there are as many predictive variables as there are cases, the fit will be perfect; (ii) the confidence of future predictions is lessened with more terms in the model; (iii) a minimal model provides some indication on the causal relationships of the included predictive variables in determining the response; (iv) monitoring costs will be reduced by measuring less environmental variables.

5.2.3.1 Predictive multiple logistic regression model

A predictive multiple logistic regression model was constructed similarly to the multiple logistic regression model constructed in Chapter 4, and the description is therefore kept short. A bi-directional stepwise model selection procedure was used [210], starting with the full model and alternately omitting and re-introducing one model component at each step. Selection stopped when no variable insertion or deletion caused a lower Akaike Information Criterion value (AIC, [212]), resulting in the model with the lowest AIC value. This methodology prevented the multiple logistic regression model from overfitting [185, 238, 249]. The residual deviance (see Eq. (4.12)), a χ^2–distributed statistic, was used to test the goodness-of-fit of the model. Models are concluded to perform satisfactorily at the α significance level if the deviance values smaller than or equal to $\chi^2(1-\alpha, \mathrm{Df})$, with Df the degrees of freedom.

5.2.3.2 Hierarchical partitioning

For the identification of important predictive variables within the multiple logistic regression context, a technique called hierarchical partitioning [238, 242, 244] was used. Hierarchical partitioning is likely to alleviate multicollinearity problems that are ignored by one-model approaches [242]. Hierarchical partitioning considers all 2^p (which is the total number of possible models using p predictive variables, including design variables) multiple logistic regression models jointly to identify the most important predictive variables. The log-likelihood (see Eq 4.9), a goodness-of-fit measure for logistic regression, is computed for each of the 2^p models. These values are partitioned so that the total independent contribution of a given predictive variable is estimated. By these means, hierarchical partitioning allows to distinguish environmental variables whose independent effect on the response variable are important, from environmental variables that have little independent effect on the response variable. More precisely, hierarchical partitioning involves the calculation of incremental improvement (increased goodness-of-fit) in models by inclusion of a given predictive variable, and these are averaged over all models in which the considered predictive variable occurs. These effects are segregated into independent effects, I (which are of interest for this study), and effects that cannot be unambiguously associated with that single predictive variable but are due to joint effects with other predictive variables, J. The output of a hierarchical partitioning analysis is a list of all predictive variables and their independent (I) and joint (J) influences on the response variable. The explicit mathematical description of hierarchical partitioning is given in Chevan and Sutherland [244]. The hier.part package [250] in R Version 2.2.1 was extended to deal with more predictive variables, and was used for this analysis.

5.2.4 Random forests

Random forests [204] is described in detail in Chapter 4, where attention was drawn on the calibration of the user-defined parameters k (number of classification trees in the random forest) and m (number of randomly taken variables to spit nodes). Algorithms 1 and 2 (see Chapter 4) explain how a random forest is constructed, and how the built-in out-of-bag error is computed, respectively. Additionally, the random forest algorithm can estimate the importance of each environmental variable by using the 'variable importance' (defined by `varimp` in code [216], as a synonym for 'predictive variable importance') measure. Defining predictive variable importances is done by looking at how much the oob error (see Algorithm 2 in Chapter 4) increases when oob data are permuted for one predictive variable while left unchanged for all others. This is done for all predictive variables. The calculation procedure for a random forest consisting of k classification trees constructed on a training data set with p predictive variables is visualized in Fig. 5.1 and given in Algorithm 5.

5.2.5 Data considerations

The data set L (Eq. (5.1)) had to be slightly adapted to support both modelling techniques.

For multiple logistic regression models (Section 5.3.3), the dependent variable (here, vegetation type, consisting of seven classes) should be binomial (0/1 = absent/present). Therefore, seven additional columns replaced the original multinomial response of L in order to include this binomial translation. The resulting data set is referred to as $L^\star$. For the best predictive multiple logistic regression model (Subsection 5.3.3.1), the data set $L^\star$ was randomly and uniformly split into three parts $L_1^\star$, $L_2^\star$ and $L_3^\star$. In 3–fold cross-validation (see Chapter 4, Algorithm 3, $k = 3$), two subsets were combined as training set, while the third was used as test set. The following abbreviations are used further on:

- MLR$_{12}$ = MLR model constructed with training set $L_1^\star \cup L_2^\star (= L_{12}^\star)$,
- MLR$_{13}$ = MLR model constructed with training set $L_1^\star \cup L_3^\star (= L_{13}^\star)$,
- MLR$_{23}$ = MLR model constructed with training set $L_2^\star \cup L_3^\star (= L_{23}^\star)$.

Model evaluation was done on the remaining independent data subset (e.g. the model MLR$_{12}$ is evaluated on $L_3^\star$). For explanatory modelling using hierarchical partitioning (Subsection 5.3.3.2), the entire data set $L^\star$ was used. Since its objective was the identification of important variables rather than the optimization of prediction accuracy, no evaluation data set was needed.

For random forest models (Section 5.3.4), a translation to a binomial response was not necessary, and the original data set L was used. For predictive random forest modelling (Subsection 5.3.4.1), the entire data set L was split into three parts L_1, L_2 and L_3 (same partitions as above). In 3–fold cross-validation, two subsets were combined as training set, while the third was used as test set. The following abbreviations are used further on:

Algorithm 5: Calculating variable importance within random forests.

```
Data: training data set $X$, with $p$ predictive variables
Result: importance (mean importance and $z$-score) of each predictive
        variable
for $i$ = 1 to $k$ do
    construct classification tree $i$ (Algorithm 1);
    apply tree $i$ to the $n$ oob elements and count the number of correct
    classifications over the $n$ oob elements ($C_{i,untouched}$);
    for $j$ = 1 to $p$ do
        take the $n$ untouched oob elements;
        randomly permute the values of variable $j$ in the $n$ oob elements;
        apply tree $i$ to all the $j$ permuted oob elements;
        count the number of correct classifications ($C_{i,j-permuted}$);
        subtract the number of correct classifications of the
        variable-$j$-permuted oob elements from the number of correct
        classifications of the untouched oob elements and divide by the
        number of oob elements ($\Delta C_{i,j} = (C_{i,untouched} - C_{i,j-permuted})/n$);
    end
end
for $j$ = 1 to $p$ do
    calculate the mean $\Delta C_{i,j}$ over all $k$ trees ($\overline{\Delta C_j} = \sum_{i=1}^{k} \Delta C_{i,j}/k$);
    refer to $\overline{\Delta C_j} \times 100$ as the ‘mean importance score’ of predictive variable
    $j$ [the value is positive when $C_{i,untouched} > C_{i,j-permuted}$ and negative
    when $C_{i,untouched} < C_{i,j-permuted}$; mean importance scores have high
    values when the classification error increases by permuting the values of
    predictive variable $j$];
    divide $\overline{\Delta C_j}$ by the standard error (se) to obtain a $z$–score for predictive
    variable $j$, and assign a significance level assuming normality [since
    correlations of the $\Delta C_{i,j}$ scores are generally low within the $j = 1$ to $p$
    groups, standard errors can be calculated for each of the $j$ groups of $k$
    $\Delta C_{i,j}$ scores] ;
end
```

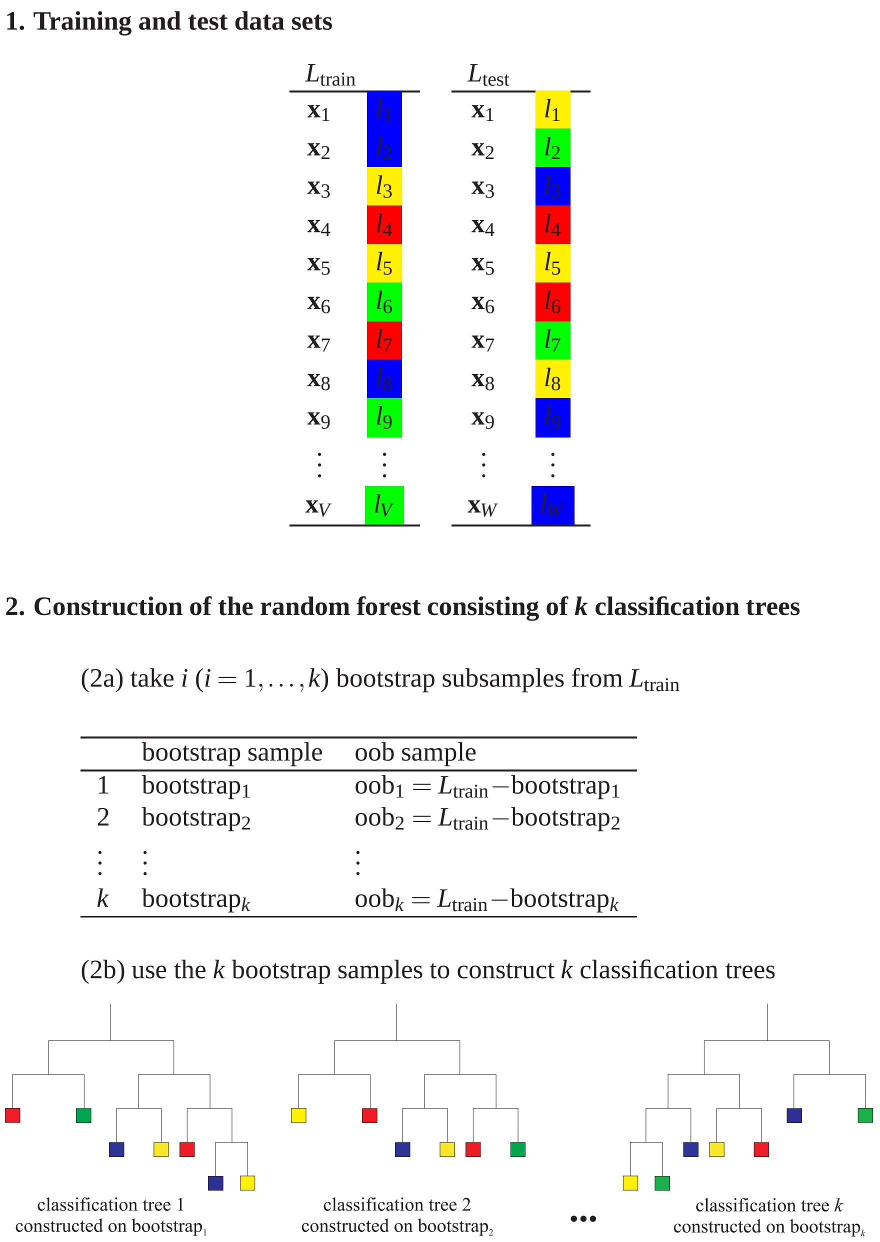
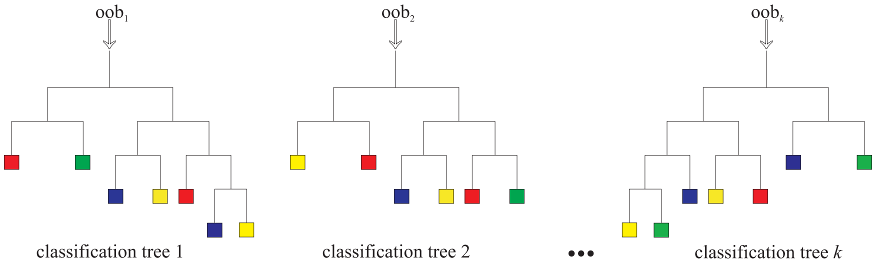

1. Training and test data sets
L_{train}
$\mathbf{x}_1$ l_1
$\mathbf{x}_2$ l_2
$\mathbf{x}_3$ l_3
$\mathbf{x}_4$ l_4
$\mathbf{x}_5$ l_5
$\mathbf{x}_6$ l_6
$\mathbf{x}_7$ l_7
$\mathbf{x}_8$ l_8
$\mathbf{x}_9$ l_9
$\mathbf{x}_V$ l_V
L_{test}
$\mathbf{x}_1$ l_1
$\mathbf{x}_2$ l_2
$\mathbf{x}_3$ l_3
$\mathbf{x}_4$ l_4
$\mathbf{x}_5$ l_5
$\mathbf{x}_6$ l_6
$\mathbf{x}_7$ l_7
$\mathbf{x}_8$ l_8
$\mathbf{x}_9$ l_9
$\mathbf{x}_W$ l_W
2. Construction of the random forest consisting of k classification trees
(2a) take i ($i = 1, \dots, k$) bootstrap subsamples from L_{train}
bootstrap sample
oob sample
1 bootstrap_1 $\text{oob}_1 = L_{\text{train}} - \text{bootstrap}_1$
2 bootstrap_2 $\text{oob}_2 = L_{\text{train}} - \text{bootstrap}_2$
k bootstrap_k $\text{oob}_k = L_{\text{train}} - \text{bootstrap}_k$
(2b) use the k bootstrap samples to construct k classification trees
classification tree 1
constructed on bootstrap_1
classification tree 2
constructed on bootstrap_2
•••
classification tree k
constructed on bootstrap_k

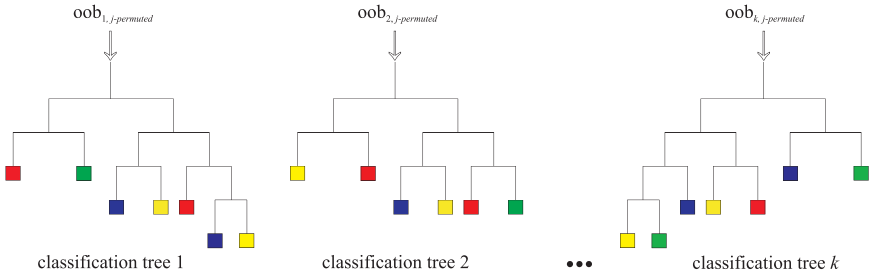
FIGURE 5.1 – Schematic overview of the construction of the random forest model and the determination of important variables.

(2c) apply classification tree$_i$ to oob$_i$ ($i = 1, \ldots, k$), and calculate oob error (oob$_{untouched}$)



3. calculate the variable importance of variable *k* (*j*=*1*, . . . , *p*)

(3a) permute the values of variable j in all oob samples (oob$_{j-permuted}$)
(3b) apply classification tree$_i$ to oob$_i$ ($i = 1, \ldots, k$)


oob$_{1, j\text{-}permuted}$
oob$_{2, j\text{-}permuted}$
oob$_{k, j\text{-}permuted}$
classification tree 1
classification tree 2
•••
classification tree k

(3c) calculate oob$_{j-permuted}$ error

4. compare the oob$_{j\text{-}permuted}$ error with the oob$_{untouched}$ error to assess variable importance

FIGURE 5.1 – *continued. . .*

- RF$_{12}$ = random forest constructed with training set $L_1 \cup L_2 (= L_{12})$,
- RF$_{13}$ = random forest constructed with training set $L_1 \cup L_3 (= L_{13})$,
- RF$_{23}$ = random forest constructed with training set $L_2 \cup L_3 (= L_{23})$.

For the identification of important variables (Subsection 5.3.4.2), however, both L and $L^\star$ were used, the latter to allow for identification of important variables for the seven vegetation types independently and to allow for comparison with results from hierarchical partitioning.

5.2.6 Model evaluation statistics

Correlations were calculated by means of the non-parametric Kendall rank correlation (Kendall's tau, τ) provided in most statistical software packages. Kendall's τ takes values between -1 and $+1$, with a positive correlation indicating that the ranks of both variables increase together, while a negative correlation indicates that as the rank of one variable increases, the other one decreases.

Model performances were evaluated in two ways. Cohen's κ test ([220], see Eq. (4.20)) was used to evaluate differences between observations and predictions for all N instances. The McNemar test [222] was used to compare error rates of two models, and is described in Section 4.5 (Eq. (4.21)).

5.3 Results

5.3.1 Data inspection

Intercorrelation amongst quantitative variables has been reported to weaken analysis using different techniques including regression [238] and ordination [193]. All 15 continuous variables were tested on normality. None was normally distributed, therefore a non-parametric correlation analysis was performed using Kendall's τ. The majority of variable pairs showed significant correlation at the 0.01 significance level (Table 5.2). Particularly high positive correlations ($\tau > 0.5$) were observed for the variable pairs: AGD – Max, AGD – Min, Max – SOM and pH – Mg^{2+}. The interpretation of the first two variable pairs is straightforward. The higher the maximal and minimal groundwater depth, the higher the average groundwater depth will be (see Fig. 3.3). Furthermore, the organic matter content is high for grid cells with high maximal groundwater depths (shallow groundwater table) due to reduced decomposition rates (see Subsection 3.3.4). Strong negative correlations ($\tau < 0.5$) were observed between Max – Ampli and Ampli – SOM. Grid cells with high maximum groundwater depth have small groundwater depth amplitudes, and the soil of grid cells with small groundwater depth amplitudes are high in organic matter content.

5.3.2 Ordinations

DCA ordination was performed, and summary statistics are given in Table 5.3. The length of gradient (expressed as standard deviation), which is a measure of how unimodal the species response is along an ordination axes, exceeds 2, approving the use of unimodal ordination models [192]. Eigenvalues of the three first

TABLE 5.2 – Kendall τ correlations between quantitative variables.

	AGD	Max	Min	Ampli	pH	Cl^-	Ca^{2+}	Fe_{tot}
AGD	1.000							
Max	0.737**	1.000						
Min	0.715**	0.481**	1.000					
Ampli	-0.359**	-0.609**	-0.090**	1.000**				
pH	0.036	-0.050	0.095**	0.166**	1.000			
Cl^-	0.197**	0.188**	0.218**	-0.092**	0.234**	1.000		
Ca^{2+}	-0.055	-0.053	-0.049	0.076**	0.400**	0.233**	1.000	
Fe_{tot}	0.058*	0.194**	-0.135**	-0.373**	-0.342**	-0.264**	-0.061*	1.000
K^+	0.162**	0.325*	0.013	-0.423**	-0.063**	-0.011	-0.110**	0.018
Mg^{2+}	-0.020	-0.067**	-0.027	0.105**	0.630**	-0.061*	0.374**	-0.071*
NO_3^-–N	-0.003	-0.046	0.150**	0.145**	0.046	0.093	-0.114**	-0.375**
NH_4^+–N	-0.339**	-0.338**	-0.328**	0.215**	-0.007**	-0.356**	0.169**	0.185**
$H_2PO_4^-$	0.244**	0.222*	0.163**	-0.210**	0.063*	0.014	-0.273*	-0.041
SO_4^{2-}	0.234**	0.280**	0.205**	-0.253**	-0.172	0.433**	0.032	-0.013
SOM	0.387**	0.517**	0.164**	-0.577**	-0.042	0.188**	0.017	0.340**

continued…

	K^+	Mg^{2+}	NO_3^-–N	NH_4^+–N	$H_2PO_4^-$	SO_4^{2-}	SOM
K^+	1.000						
Mg^{2+}	-0.018	1.000					
NO_3^-–N	-0.043	-0.212**	1.000				
NH_4^+–N	-0.314**	0.132**	-0.099**	1.000			
$H_2PO_4^-$	0.063*	0.109**	-0.097**	-0.165**	1.000		
SO_4^{2-}	0.174**	-0.410**	0.138	-0.410**	-0.130**	1.000	
SOM	0.254**	-0.021	-0.167**	-0.160**	0.155**	0.201**	1.000

* correlation is significant at the 0.05 level; ** correlation is significant at the 0.01 level.

DCA ordination axes (Table 5.3) cannot be interpreted as proportions of variance explained since the process of rescaling and detrending destroys the correspondence between the eigenvalues and the structure along the axes. Therefore, the variance explained was investigated by a *post hoc* calculation of the coefficient of determination (r^2) between distances in the ordination space and distances in the original space. The Euclidean distance was used as distance measure of the ordination space, while the relative Euclidean distance was selected as distance measure of the original space (Table 5.3). High and moderate correlations between the original and ordination space were found for the first two axes, 0.476 and 0.198, respectively. Along the third axis, correlations were negligible.

Fig. 5.2 jointly plots the plant species and vegetation types positioned in the DCA ordination space, together with the environmental variables (labels of K^+, NO_3^-–N and no management are deleted for clearness). Kendall correlations (τ) of the environmental variables with ordination axes were calculated, and given for the two main ordinations axes in Table 5.1. The first axis was highly correlated with groundwater quantity and dynamics, predominantly average groundwater depth, maximal groundwater depth and amplitude of the groundwater depth, soil type and soil organic matter content, and the management variables, predominatly yearly mowing. As can be concluded from Fig. 5.2 and Table 5.1: the *Arrhenatherion elatioris* grassland community thrives on the drier, loam soils with low organic matter content of the Doode Bemde with a yearly mowing management. These conditions

TABLE 5.3 – Summary statistics for the DCA and CCA ordinations.

	Axis 1	Axis 2	Axis 3
DCA			
Length of gradient	5.99	5.55	3.57
Eigenvalue	0.557	0.339	0.174
r^2	0.476	0.198	0.060
Cumulative r^2	0.476	0.673	0.733
CCA			
Eigenvalue	0.501	0.270	0.208
Variance explained [%]	10.6	5.7	4.4
Cumulative variance explained [%]	10.6	16.3	20.7
r^2	0.451	0.111	-0.023
Cumulative r^2	0.451	0.562	0.585

r^2 is the coefficient of determination.

are typically situated on the southwestern levee. *Filipendulion* and *Calthion palustris* are found in the transitional zone between levee and floodplain depression (see Fig. 3.2). Both vegetation types are clearly distinguishable along the second axis, with *Calthion palustris* occurring in the somewhat wetter, more peaty areas. The groundwater table in the *Filipendulion* areas shows higher fluctuations. Another difference between these vegetation types is the mowing frequency. The majority of the *Calthion palustris* sites are yearly mown, while most of the *Filipendulion* sites are cyclically mown. Still further toward the floodplain depression *Carici elongetae – Alnetum glutinosae* and the tall sedge vegetation *Magnocaricion* appear on wet, peaty soils. Sedges are gradually replaced by *Phragmites australis* and the *Magnocaricion* vegetation type changes in *Magnocaricion* with *Phragmites* in the central part of the floodplain. On the wettest part of the eastern riparian zone, a *Phragmitetea* vegetation belt occurs. The most important chemical groundwater variables are SO_4^{2-}, Cl^-, NH_4^+–N, and Fe_{tot}, with high concentrations of SO_4^{2-} and Fe_{tot} in the wetter areas, where groundwater seepage and capillary rise increase the supply of these hydro-chemical compounds (see Subsection refgwqual), and high Cl^- concentrations in the moderately wet northeastern areas. High concentrations of NH_4^+–N typically occur in the dryer, loamy areas with *Arrhenatherion elatioris*, and can possibly be attributed to a higher rate of biological mineralization of organic nitrogen (e.g. in plant litter) within this dryer area [2].

Several variables were deleted for the CCA due to multicollinearity problems. The only variables related to groundwater quantity retained were average groundwater depth and amplitude of the groundwater depth (maximal and minimal groundwater depth were deleted based on their high correlations with average and amplitude of the groundwater depth, see Table 5.2). The categorical variables ‘no management’ and the soil type ‘peat’ were deleted without loss of informa-

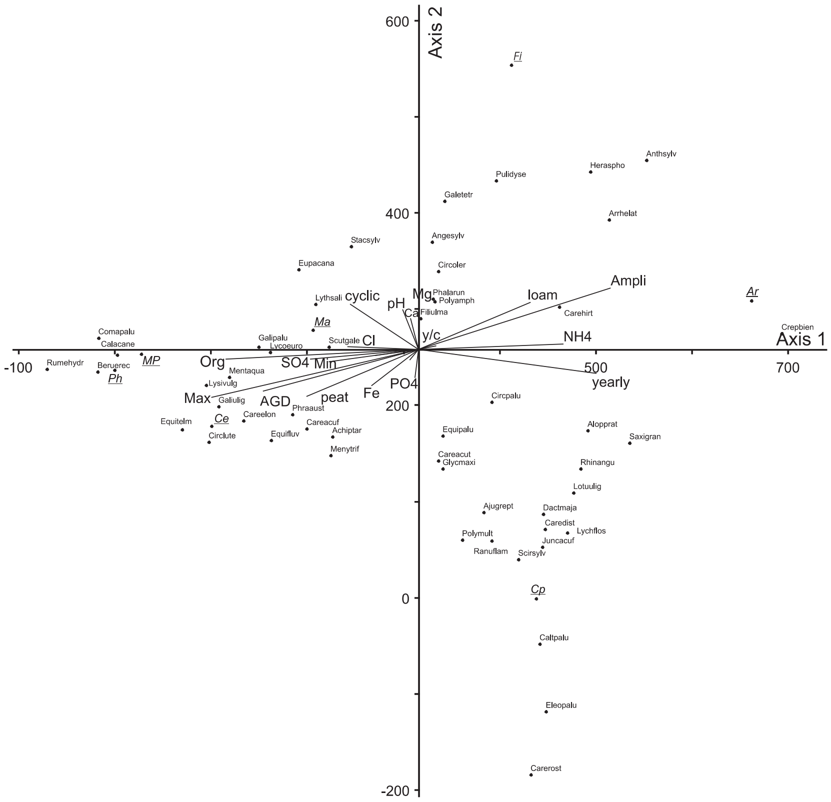


FIGURE 5.2 – Graph of the detrended correspondence analysis (DCA) showing the ordination of plant species and vegetation types along the first two ordination axes in relation to the environmental variables.
Legend: Species are abbreviated using the first four letters of the genus and species names (as given in Appendix A), vegetation types are abbreviated according to the List of Abbreviations and Acronyms, and abbreviations of the environmental variables: SOM = soil organic matter content, y/c = transition from yearly to cyclic mowing, chemical ions are given without charge.

tion on management regime and soil type. The CCA eigenvalues (Table 5.3) were a little lower than the DCA eigenvalues. The first two axes explained a limited 16.0% of the variance in data. In analogy with DCA, correlation between distances in the ordination space and distances in the original space were determined. Once again, the relative Euclidean distance was used as distance measure in the original space. r^2 for CCA and the effectiveness of the CCA were considerably lower than what was obtained for DCA. Nevertheless, the Kendall correlation (τ) between the species scores on the DCA and CCA axes equaled 0.881 ($p < 0.01$) and 0.534 ($p < 0.01$) for the first and second axis, respectively. There is a strong positive correlation between the first axes of both ordination methods. The CCA ordination is shown in Fig. 5.3, and Table 5.1 gives Kendall's correlation values for the environmental variables with the two main CCA axes. In accordance with the DCA ordination, average groundwater depth, soil type, organic matter content, mowing management regime and the chemical variables NH_4^+–N, SO_4^{2-} and Cl^- did explain most variance along the first CCA axis. Important variables along the second CCA axis included Ca^{2+} and $H_2PO_4^-$.

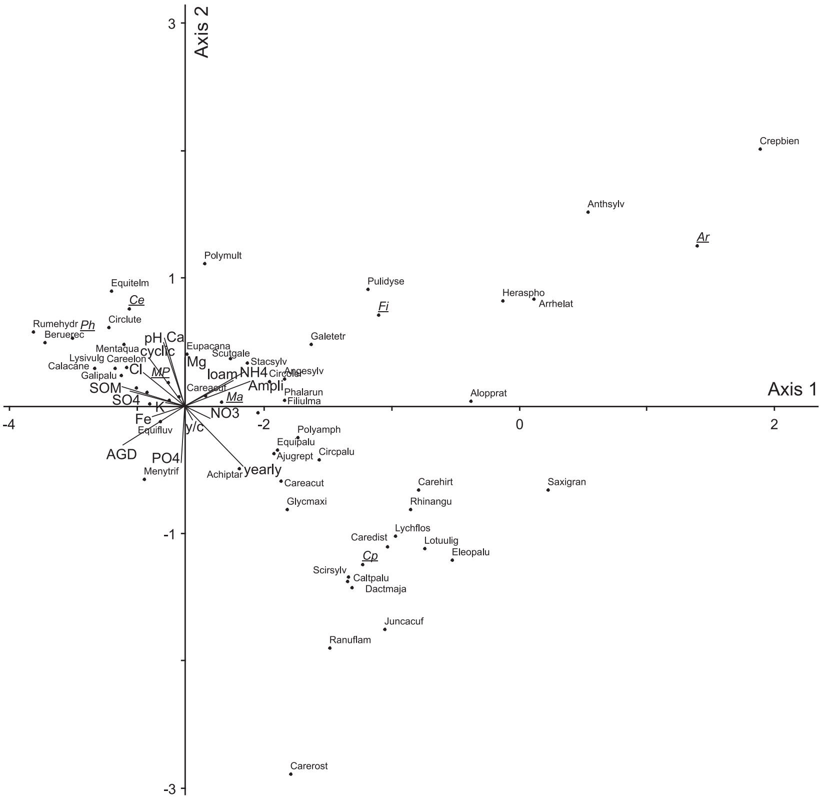


FIGURE 5.3 – Graph of the canonical correspondence analysis showing the position of plant species and vegetation types in relation to the environmental variables.
Legend: Species are abbreviated using the first four letters of the genus and species names (as given in Appendix A), vegetation types are abbreviated according to the List of Abbreviations and Acronyms, and abbreviations of the environmental variables: SOM = soil organic matter content, y/c = transition from yearly to cyclic mowing, chemical ions are given without charge.

In accordance with De Becker et al. [81] the ordinations indicated the groundwater describing variables, soil type and organic matter content and the yearly mowing management regime as the most important variables at the Doode Bemde.

Together with the groundwater quality variables SO_4^{2-}, Cl^- and NH_4^+–N these were the environmental variables explaining most of the spatial variance of plant species and vegetation type occurrences at the Doode Bemde. NO_3^-–N, K^+, Mg^{2+} and the transition from yearly to cyclic mowing management are concluded to be less important based on both ordinations.

5.3.3 Multiple logistic regression model

5.3.3.1 Predictive multiple logistic regression model

In order to use 3–fold cross-validation, three models (MLR_{12}, MLR_{13}, MLR_{23}) had to be constructed on the three training sets $L^\star_{12}$, $L^\star_{13}$ and $L^\star_{23}$. Additionally, logistic regression required to model the seven vegetation types separately (seven models). Consequently, a total of 21 multiple logistic regression models were constructed, one on each of the three training data sets for each of the seven vegetation types. Since high correlations between average groundwater depth and maximal and minimal groundwater depth tended to weaken the multiple regression models, it was decided to include only the average groundwater depth and amplitude of the groundwater table as groundwater quantity variables. The goodness-of-fit of all 21 models was summarized by the residual deviance (D_{resid}, Eq. (4.12)). Based on the results reported in Table 5.4, it can be concluded that all multiple logistic regression models do fit satisfactorily at the 0.01 significance level. Nevertheless, the models used for *Phragmitetalia* and *Arrhenatherion elatioris* (lower residual deviance values) are better than those for *Calthion palustris* (higher residual deviance values).

The models were applied to their corresponding independent test data set $L^\star_1$, $L^\star_2$ or $L^\star_3$. Each 20 m by 20 m grid cell at the Doode Bemde was assigned seven probabilities, representing the probability of occurrence of the seven different vegetation types. The vegetation type with the highest probability of occurrence was concluded to be the predicted vegetation type for the grid cell under consideration. Results are visualized in Fig. 5.4(a). There were 359 cells correctly predicted (69.2%), on a total of 519 grid cells. Cohen's κ test was used to evaluate differences between observations and predictions. A κ value of 0.633 was found: there is a substantial agreement between observations and predictions ($p < 0.001$).

5.3.3.2 Explanatory modelling with hierarchical partitioning

Each predictive variable explains a certain amount of the spatial vegetation pattern at the Doode Bemde. For each vegetation type separately, a logistic regression model was constructed, and the model goodness-of-fit was assessed, together with the independent contribution I of each predictive variable to the explanation of the spatial distribution of that vegetation type. I–values are graphically presented in Fig. 5.5. For some vegetation types, a clear distinction between the most important

TABLE 5.4 – Goodness-of-fit of the multiple logistic models. Residual deviances ($D_{\text{resid.}}$) and Akaike's information criterion (AIC) for the different vegetation types.

Vegetation type	Df	D_{resid}	AIC
MLR$_{12}$			
Arrhenatherion elatioris	334	34.7*	59.0
Calthion palustris	334	193.8*	218.0
Carici elongetae – Alnetum glutinosae	337	80.2*	98.2
Filipendulion	335	134.2*	156.2
Phragmitetalia	332	38.1*	67.4
Magnocaricion with *Phragmites*	335	132.6*	154.6
Magnocaricion	333	154.9*	180.9
MLR$_{13}$			
Arrhenatherion elatioris	339	51.4*	56.4
Calthion palustris	334	168.8*	192.8
Carici elongetae – Alnetum glutinosae	336	67.5*	87.8
Filipendulion	338	122.3*	138.3
Phragmitetalia	337	29.6*	47.6
Magnocaricion with *Phragmites*	335	120.2*	142.2
Magnocaricion	332	142.5*	170.5
MLR$_{23}$			
Arrhenatherion elatioris	338	66.3*	82.7
Calthion palustris	334	190.9*	214.9
Carici elongetae – Alnetum glutinosae	334	60.9*	84.9
Filipendulion	333	142.8*	168.8
Phragmitetalia	334	33.2*	57.7
Magnocaricion with *Phragmites*	328	131.6*	178.7
Magnocaricion	339	138.9*	152.9

Df = degrees of freedom, D_{resid} = residual deviance, AIC = Akaike's information criterion. All models showed a significant goodness-of-fit at the 0.01 level (*).

predictive variable and the other ones can be observed (e.g. management regime for *Calthion palustris*), while for other vegetation types a group of important predictive variables is determined (e.g. average groundwater depth, management, pH and Fe_{tot} for *Magnocaricion*). The variables, ranked according to their independent contribution to the distribution model, do differ between the seven vegetation types.

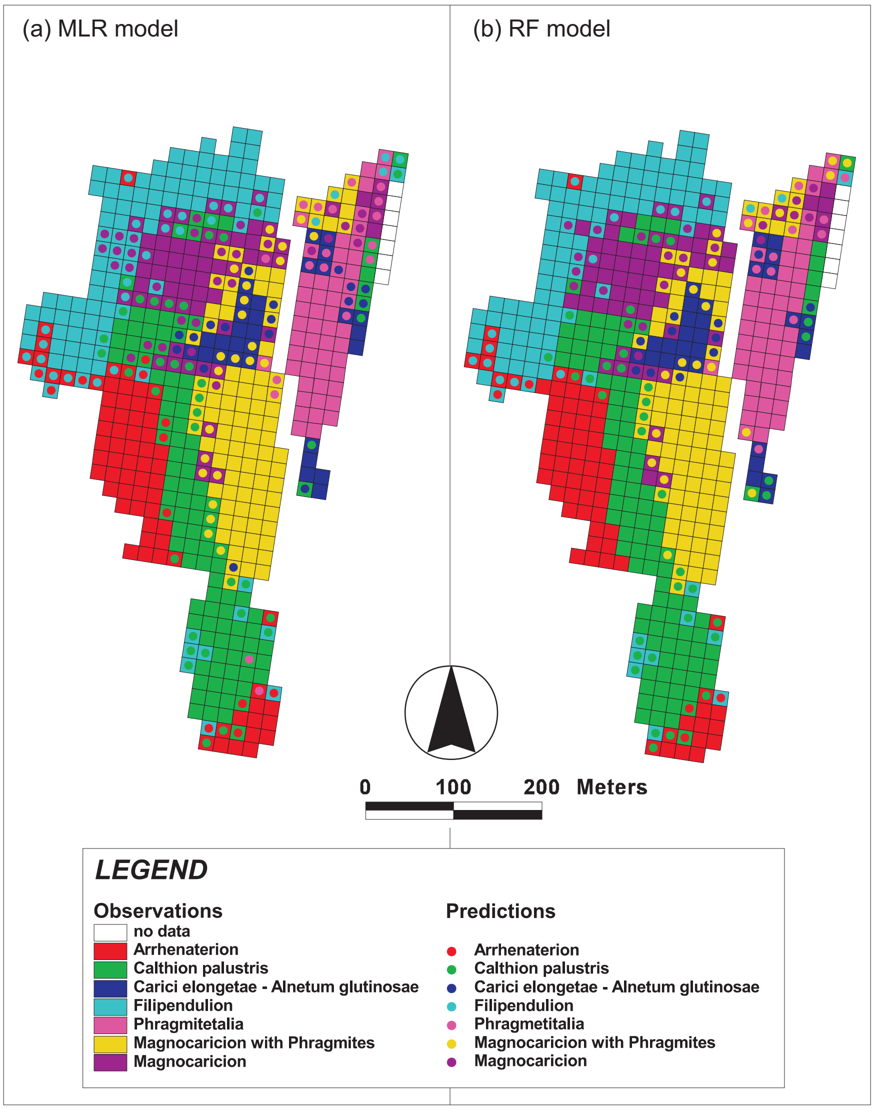


FIGURE 5.4 – Spatially distributed vegetation types at Doode Bemde. Observations overlaid by predictions made by the logistic regression models (a). Observations overlaid by predictions made by the random forest models (b).

5.3.4 Random forest model

5.3.4.1 Construction and evaluation of the best predictive random forest model

The random forest technique has two important user-defined parameters that should be optimized for accurate model results: the number of trees (k) and the

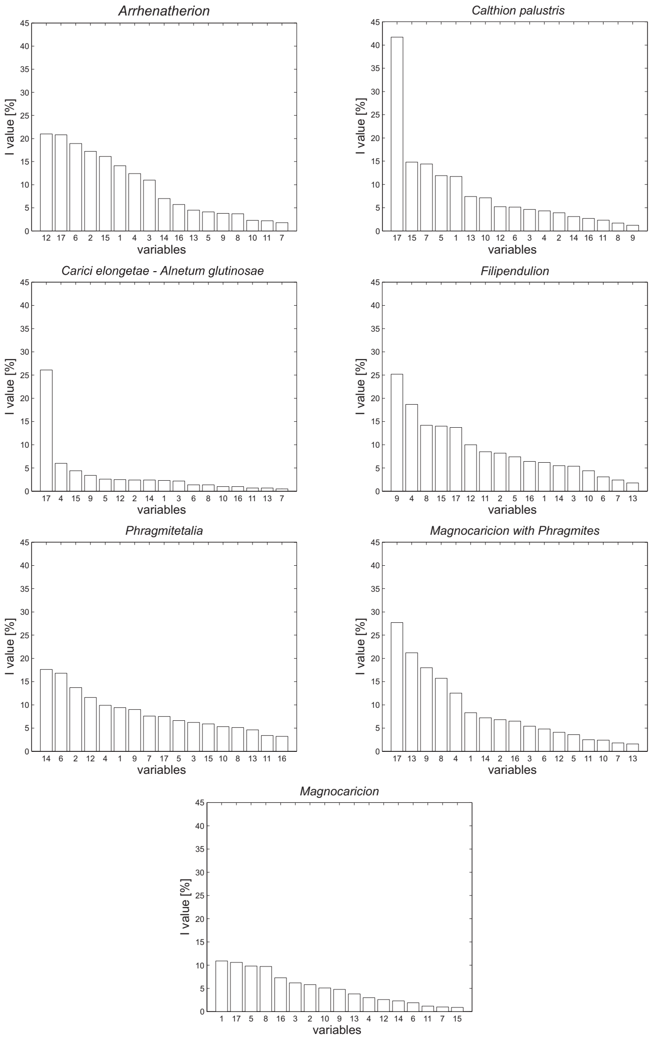


FIGURE 5.5 – Independent individual contributions (I–values, [%]) of the variables as determined with hierarchical partitioning. Lower values indicate lower independent importance.
Legend: The numbers on the x–axis correspond to variables: average groundwater depth (1), minimal groundwater depth (2), maximal groundwater depth (3), amplitude of the groundwater depth (4), pH (5), Cl^- (6), Ca^{2+} (7), Fe_{tot} (8), K^+ (9), Mg^{2+} (10), NO_3^-–N (11), NH_4^+–N (12), $H_2PO_4^-$ (13), SO_4^{2-} (14), soil organic matter content (15), soil type (16) and management regime (17).

number of randomly selected predictive variables to split the nodes (*m*) (see Subsection 4.4.2). Oob error, which was proven to be an unbiased estimator of the

Table 5.5 – Out-of-bag (oob) error values and test set error values for RF$_{12}$, RF$_{13}$ and RF$_{23}$ constructed with several values of *m* (*m* is the number of randomly sampled variables to split the nodes).

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 15$
[a]RF$_{12}$	35.84	23.41	24.57	23.99	23.41	25.43
[b]L_3	32.95	21.97	22.54	23.12	23.70	23.12
[a]RF$_{13}$	30.35	21.10	20.23	19.36	19.94	21.39
[b]L_2	34.10	28.32	24.28	26.59	26.01	27.75
[a]RF$_{23}$	40.17	23.70	22.54	22.54	24.28	23.99
[b]L_1	45.09	22.54	21.39	22.54	21.39	23.70

[a] = oob error, [b] = test set error. L_1, L_2 and L_3 are cross-validation test data sets. Minimal error values are underlined.

classification error [204], was used to determine an adequate number of trees. A clear convergence of oob error was found (not shown here, but similar to Figure 4.6), and $k = 1000$ was used for the construction of all the random forest submodels. Breiman and Cutler [218] state that the range of optimal values of *m* is usually quite wide, and often $\sqrt{\text{number of predictive variables}}$ is used as value for *m*. According to this rule of thumb, in this study, where 17 variables were used, the optimal value of *m* should be around 4. Three random forest models were constructed for different values of *m*. Two error measures were used for optimal *m* definition: (i) oob error, and (ii) test set error, representing the proportion of incorrect classifications, which is computed by applying the random forest models to their respective test data sets during the random forest building process. Resulting oob error values and test set error values are tabulated in Table 5.5. The values did not differ greatly, but minimal values could be observed for different values of *m*. Since an accurate classification of the test elements was the goal, it was decided to use the values of *m* for which the test set error was minimal and to take $m = 2$ for the construction of RF$_{12}$, and $m = 3$ for RF$_{13}$ and RF$_{23}$. These values are lower than 4, which was determined by the rule of thumb.

Based on the optimal values for the user-defined model parameters *k* and *m*, RF$_{12}$, RF$_{13}$ and RF$_{23}$ were constructed. Cross-validation resulted in an independent prediction for each of the 519 grid cells at the Doode Bemde. Results are visualized in Fig. 5.4(b). On a total of 519 grid cells, the model made 402 correct classifications (77.5%) and 117 misclassifications (22.5%). Misclassifications were mostly made on the transition zone between adjacent vegetation types as well as for small vegetation patches (e.g. one grid cell surrounded by a different vegetation type). A κ value of 0.731 was found and indicated a substantial agreement between observations and predictions ($p < 0.001$).

5.3.4.2 Explanatory modelling using the 'variable importance' measure

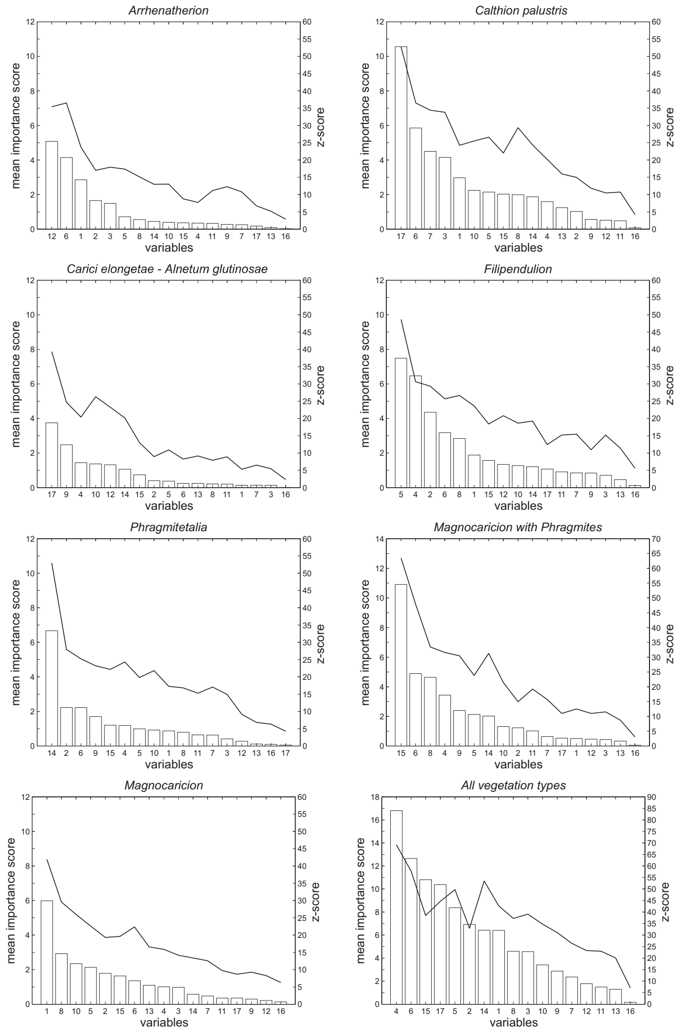
To allow for comparison of the important variables as assessed with hierarchical partitioning and the 'variable importance' measure, seven new random forest models were constructed based on the data set $L^{\star}$, one random forest model for each vegetation type. The user-defined parameter m was set to 17, so that the random forest had to choose the best predictive variable to split the nodes amongst all 17 possibilities. However, by doing so, the only random process during model construction was the use of random bootstrap samples, and therefore the correlations between the different classifiers in the random forest are likely to increase. Resulting mean importance scores ($\overline{\Delta C_j} \times 100$) and z–scores ($\overline{\Delta C_j}/\mathrm{se}$, with se the standard error) from the 'variable importance' measure used during the construction of these models are plotted in Fig. 5.6. Variables are ranked from high to low mean importance scores, z–scores are overlaid. Mean importance scores and z–scores generally showed (more or less) the same tendency. However, since correlations between the classifiers in the random forest were likely to be significant (m was set to its maximal value 17 to exclude randomness from predictive variable selection), preference was given to the mean importance scores to determine important predictive variables for spatial distribution modelling. Differences in importance scores could be observed between the different vegetation types, and *post hoc* data inspection was needed to determine whether high or low values of important variables were associated with vegetation type occurrences. As an example, the distribution of *Calthion palustris* in the study area is clearly related with management regime, which management regime had to be elucidated by a posterior data inspection. In this case *Calthion palustris* is associated with yearly mowing.

Based on data set L, the variables were ranked according to their mean importance for all vegetation types together (Fig. 5.6). The amplitude of the groundwater table is the most important variable for the vegetation type distribution at the Doode Bemde, followed by Cl^-, soil organic matter content and management regime. Soil type and several chemical groundwater variables ($\mathrm{H_2PO_4^-}$, $\mathrm{NO_3^-}$–N and $\mathrm{NH_4^+}$–N) are clearly the least important variables for vegetation type distribution modelling with the random forest model at the Doode Bemde.

5.3.5 Predictive random forest modelling on reduced data subsets

5.3.5.1 Leave-one-variable-out

It was believed that important predictive variables would have major influences on classification accuracy. Exclusion of a predictive variable would consequently result in an increase in oob error, proportional to predictive variable importance.


Arrhenatherion

Calthion palustris

Carici elongetae - Alnetum glutinosae

Filipendulion

Phragmitetalia

Magnocaricion with Phragmites

Magnocaricion

All vegetation types

FIGURE 5.6 – Mean importance scores (bars) and *z*–scores (lines) for all 17 variables, for each vegetation type separately and all vegetation types together, as determined with the random forests 'variable importance' measure. Lower values indicate lower importance.
Legend: The numbers on the x–axis correspond to variables: average groundwater depth (1), minimal groundwater depth (2), maximal groundwater depth (3), amplitude of the groundwater depth (4), pH (5), Cl^- (6), Ca^{2+} (7), Fe_{tot} (8), K^+ (9), Mg^{2+} (10), NO_3^-–N (11), NH_4^+–N (12), $H_2PO_4^-$ (13), SO_4^{2-} (14), soil organic matter content (15), soil type (16) and management regime (17). Notice different scales for subplot 6 and 8.

Therefore data set *L* was redesigned into 17 new data sets containing all but one predictive variable (16 predictive variables in total). 17 random forest models were constructed using these data sets. An increase in oob error proportional with the predictive variable importances was expected. However, no such increase was observed. A mean oob error value of 21.20%, within the range 20.23% – 22.35%, was found. 20.04% of the oob cases were misclassified when soil type or minimal water table depth were excluded from the data set, and 22.35% if Mg^{2+} was excluded. The random forest technique was concluded to be a strong classifier, able to construct models with comparable accuracy levels when only one predictive variable, irrespective of its importance, was excluded.

5.3.5.2 Gradually decreasing model complexity

Results from explanatory modelling using the ‘variable importance’ measure were used to construct data sets with a gradually decreasing number of predictive variables. 17 random forest models were constructed using backward elimination of the least important predictive variables (the ranking of the predictive variables can be seen in the lower right panel of Fig. 5.6) of data set L. The oob error was used to get an estimation of the model classification error. As can be seen in Fig. 5.7, the oob error ranged around 22% for the random forest models constructed on all 17 predictive variables, for the model containing 6 predictive variables, and for all models with a complexity in between. The models containing less than the 6 most important predictive variables, however, showed a sharp increase in oob error. Remarkably high oob values were found for the models with only the three, two and one (oob error $\approx$ 72%) most important predictive variables included. This allows to conclude that at least 6 important predictive variables are needed for accurate vegetation distribution modelling in this study. These predictive variables are amplitude of the groundwater depth, Cl^-, organic matter content, management, pH and minimal groundwater depth.

5.3.5.3 Predictive modelling using the selected subset

Three reduced predictive random forest models were constructed on the data subset containing the six most important predictive variables: amplitude, Cl^-, organic matter content, management, pH and minimal groundwater depth. 3–fold cross-validation resulted in independent vegetation type predictions for each of the 519 grid cells at the Doode Bemde. 384 predictions were correct (74%), 135 wrong (26%). A Cohen’s κ value of 0.691 was found: there is a substantial agreement between observations and predictions ($p < 0.001$).

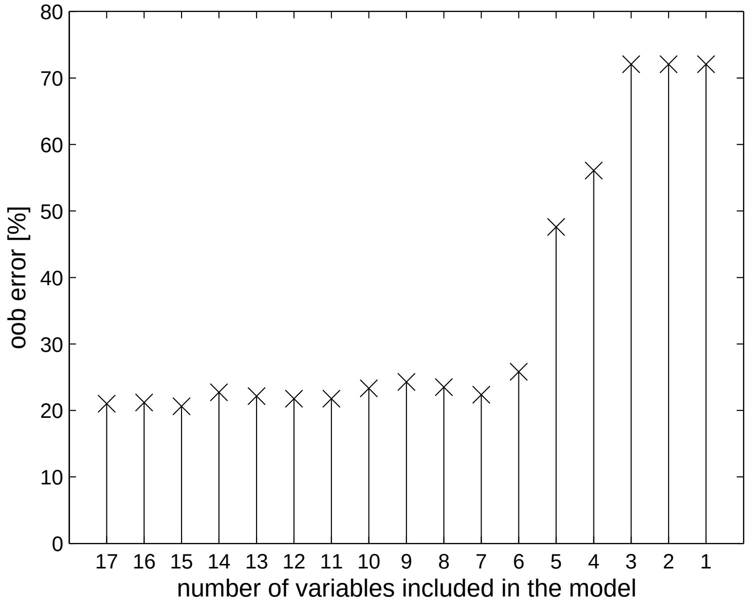


FIGURE 5.7 – Out-of-bag (oob) error of models with decreasing number of predictive variables. The order of deletion is taken from results for all vegetation types, as presented in subplot 8 of Fig. 5.6.
Legend: The numbers on the X-axis correspond to the number of predictive variables included in the model e.g. 1= only the most important predictive variable (amplitude) included; 2= two most important predictive variables (amplitude and Cl^- included), etc.

5.4 Statistical model evaluation and discussion

5.4.1 Predictive modelling

The performance of the predictive logistic regression model was compared with that of the predictive random forest model. The McNemar test was used to test the following null hypothesis: the two models have the same classification error rate. The value of the McNemar test statistic was 24.2 and the null hypothesis could be rejected ($p < 0.001$). Classification error was significantly lower for the random forest model. In accordance with Chapter 4 and [131], the random forest model could be concluded to be more accurate than the multiple logistic regression model.

TABLE 5.6 – Kendall τ correlations between results of hierarchical partitioning (*I*–values) and ‘variable importance’ (mean importance scores).

Vegetation type	τ value
Arrhenatherion elatioris	0.303
Calthion palustris	0.441^{*}
Carici elongetae – Alnetum glutinosae	0.544^{**}
Filipendulion	0.221
Phragmitetalia	0.471^{**}
Magnocaricion with *Phragmites*	-0.221
Magnocaricion	-0.074

* correlation is significant at the 0.05 level;
** correlation is significant at the 0.01 level.

5.4.2 Explanatory modelling

The logistic regression and random forest modelling approach made use of different techniques to identify important environmental variables, i.e. hierarchical partitioning [238, 242, 244] and ‘variable importance’ [204], respectively. Both techniques indicated the same predictive variable as the most important one for five out of seven different vegetation types in distribution modelling. However, to allow for comparison of the results of both techniques, the non-parametric Kendall’s τ was used. For the seven different vegetation types separately, the *I*–values (as determined by hierarchical partitioning) and mean importance scores (as determined by the ‘variable importance’ measure) of all predictive variables were ranked. Kendall τ correlations were calculated for these ranked predictive variables, and tested for significance (Table 5.6). For *Calthion palustris*, *Carici elongetae – Alnetum glutinosae* and *Phragmitetalia* the predictive variable importance ranking as determined with hierarchical partitioning showed significant similarities with the ranking as determined by ‘variable importance’ at the 0.05 significance level. But for the other vegetation types ranked predictive variables were not significantly correlated. These statistics show that different predictive variables have different effects on the goodness-of-fit of distribution models for most of the vegetation types depending on whether logistic regression or the random forest technique is used.

Major advantages of the random forest ‘variable importance’ measure were computation time and memory requirements. Hierarchical partitioning of $2^{19} = 524288$ multiple logistic regression models took approximately 8 hours (on a SGI Origin 300), while random forest modelling finished within half a minute. Furthermore, the use of categorical variables, which are frequently used in applied ecology to simplify data collection [251], is more complicated using hierarchical partitioning since translation to dummy variables is needed. Thirdly, the necessity to calculate predictive variable importances for all vegetation types separately using hierarchical partitioning was experienced as a shortcoming, since general descriptions about the major environmental gradients and related vegetation type distributions in the entire study area could not be made.

A comparison of the DCA ordination (Fig. 5.2) and the random forest variable importances was made. In general, most of the important variables coincided for both methodologies. Spatial differences in average groundwater depth and amplitude, management, soil organic matter content and pH, Cl^- and SO_4^{2-} concentrations explained most of the spatial vegetation distribution according to both methodologies. However, unlike ordination, the ‘variable importance’ measure has no direction nor sense. The ecological position of the vegetation types on the environmental gradients could be seen directly. A numerical or visual *post hoc* determination was therefore needed. Additionally, the NH_4^+–N concentration, which is an important gradient along the first DCA-axis and is associated with the distribution *Arrhenatherion elatioris* at the Doode Bemde, was not found to be important in random forest modelling of the entire data set. Nevertheless, the random forest distribution model for *Arrhenatherion elatioris* indicated NH_4^+–N as the most important variable.

Furthermore, the categorical variable soil type (peat, loam) was identified as an important environmental variable by the ordination analysis. In the random forest models, however, this variable was of very little importance. The explanation lies in the categorical nature of this variable. Since there are only two soil type classes (loam and peat), the probability that an oob element belongs to the same soil type before and after permutation is rather large. This probability is even higher with prevailing classes within a categorical variable (as observed for soil type at the Doode Bemde: 191 grid cells have loam soils, 328 grid cells have peat soils). This effect is likely to diminish when a categorical variable consists of more classes with similar numbers of elements. It is reasonable to conclude that the ‘variable importance’ measure is not suitable for handling categorical variables with a small number of categories.

This conclusion was compared with literature, and could be extended since Strobl et al. [252] proved that the ‘variable importance’ measure is not only affected by the number of categories of categorical predictive variables, but also by the scale of measurement of continuous predictive variables, which are both no direct indicators of importance. The reasons why the random forest ‘variable importance’ measure is biased is twofold: (i) there is bias in variable selection in the individual classification trees, and (ii) there is bias induced by bootstrap sampling [232, 252, 253]. Bias in variable selection for node splitting in individual trees results from the systematic preference for predictive variables with a higher number of possible cutpoints. More possible cutpoints means a higher likelihood to produce a good split. Categorical predictive variables with more categories and continuous predictive variables with a wider value range comprise more possible cutpoints. Therefore these predictive variables are selected more frequently to split the nodes and the nodes they split tend to be situated closer to the root of each classification tree [252]. Predictive variables that appear more frequently and that are situated closer to the root of each classification trees within the random forest affect the prediction accuracy of a larger subset of out-of-bag elements, while predictive variables that appear less frequently and more toward the leafs of the classification tree within the random forest affect smaller subsets of out-of-bag elements, resulting in a biased predictive variable importance estimation. Additionally, bootstrap sampling with replacement (see Algorithm 1, Chapter 4) introduced bias into the variable importance estimate. Strobl et al. [252] explain this source of bias by considering *p*-values of χ^2-tests computed from 1000 simulated data sets. They generated four artificial categorical, independent variables with a multinomial distribution with values in $\{0,\ldots,k-1\}$, where k is 2, 4, 10 and 20, for the four data sets, respectively. The values $\{0,\ldots,k-1\}$ had equal probabilities. The response variable was sampled from a binomial distribution, and independent from the independent variables. Under the null hypothesis of independence, a range of *p*-values of the χ^2-tests from 0 to 1 were calculated, with a median *p*-value of 0.5 (as expected). However, when the same analysis was made on bootstrap samples from the four variables, the median *p*-value differed significantly from 0.5, as a clear shift toward a *p*-value of 0 was observed. The bootstrap sampling artificially induced an association between independent and dependent variable. Furthermore, the association was more pronounced for independent variables with more categories, which showed a higher deviation from the null hypothesis. The apparent association affects the ‘variable importance’ because the higher association for independent variables with more categories results in a higher selection frequency, and again, a selection closer to the root of the individual trees.

Finally, attention should be drawn to the fact that important variables, as determined by ordination, hierarchical partitioning and ‘variable importance’, are defined for the specific conditions at the Doode Bemde (within the limitations of the data set). Since the data set consists of measurement vectors assigned to a certain vegetation type as observed in the field, only realized niches (the part of all suitable habitats where species and vegetation types are not excluded due to biotic interactions [237], see Section 1.4) are included. Conclusions with regard to the fundamental ecological niche of the different vegetation types are consequently hard to make. The identification of causal environmental variables may be valid within the study area, but is only a part of the story beyond that spatial limitation.

5.4.3 Reduced random forest model

The most important variables as determined with the ‘variable importance’ measure were used to construct and cross-validate a random forest model based on a reduced variable subset. A sharp increase in oob error indicated six to be the minimal number of variables to include. The classification error was 3.5% higher for the reduced model (26.0%) compared to the full model (22.5%). The McNemar test was used to statistically compare classification accuracy between the reduced random forest model and the full random forest model constructed on the entire data set. No significant difference in model performance was found at the 0.05 significance level ($p = 0.016$), however at the 0.01 significance level the difference in model performance was significant in favor of the full random forest model. The same test statistic was used to compare classification accuracy between the reduced random forest model and the multiple logistic regression model, and a significant better performance of the reduced random forest model was found ($p = 0.002$).

Apart from the higher generalization potential of the reduced random forest model (reduced number of degrees of freedom), the inclusion of limited numbers of variables and hence monitoring efforts, has clear benefits from an economical point of view. In the particular case of the Doode Bemde, for example, a costly soil type inventory could be omitted. Moreover, in most ecohydrological modelling exercises, temporal variability is averaged out, making monitoring of highly dynamic variables in time somewhat irrelevant (due to the significance of the sampling date). Based on the identification of important variables at the Doode Bemde, less of such variables have to be monitored for constructing adequate vegetation distribution models for comparable groundwater dependent valley ecosystems in Flanders.

5.5 Conclusions

Together with the high level of biodiversity [78, 81, 254], lowland river valleys are playing an important role in the context of integrated water management including flood control and sediment transport reduction [255, 256]. The key environmental variables of these ecosystems are of primary importance for assessing the feasibility of ecosystem restoration [257].

A dichotomous structure was prevalent in this chapter: (i) the evaluation of predictive distribution models using different modelling techniques (predictive modelling), and (ii) the assessment of the possibility to identify the key environmental variables at the Doode Bemde (explanatory modelling), and the linkage between both modelling approaches, namely to evaluate the predictive ability of a reduced distribution model based on these key variables only. The following conclusions could be drawn:

1. Predictive models were constructed using multiple logistic regression and the random forest technique. An accuracy evaluation of the cross-validated data proved that random forest models do perform significantly better (in accordance with results of Chapter 4).

2. Important variables could be identified using hierarchical partitioning of the logistic regression models and ‘variable importance’ within the random forest models. Groundwater dynamics and to a lower extent management practices were the predominant variables constraining the spatial vegetation distribution. However, for certain vegetation types other environmental variables seemed important, but an ecological interpretation of these results is difficult to make.

3. A bias in estimated variable importance using random forest’s ‘variable importance’ was found for categorical variables, and similar observations were found in literature.

4. Nevertheless, the ‘variable importance’ measure could be used for subset selection. The reduced predictive random forest model, resulting from a data set only containing the six most important variables, was tested on prediction accuracy and compared with the best predictive random forest model. The accuracy of the reduced model and the full model did not differ at the 0.05 significance level. Furthermore, the reduced random forest model performed significantly more accurately at the 0.01 significance level than the multiple logistic regression model.

6

Independent model testing

6.1 Introduction

Wetland ecosystems are complex, evolving structures whose characteristics and dynamic properties depend on many interrelated links between hydrology, the physicochemical environment, anthropogenic disturbances and vegetation, and their environmental determinants (climate, geomorphology) (see Fig. 1.4). The direct effect of site hydrology on physicochemical site properties, such as soil moisture content, oxygen and nutrient availability determines the productivity and species composition of the wetland vegetation [258, 259]. Vegetation, however, is not passive to the abiotical setting, but affects site hydrology and physicochemical properties through feedback processes of which transpiration [32], soil aeration [29] and alterations in nutrient loadings [260, 261] are just some examples. These localized direct and feedback processes result in spatial and temporal distributions of the abiotical variables [49]. Together with intraspecific, interspecific and anthropogenic interactions these distributed abiotical variables constrain plant species occurrences, resulting in vegetation patterns.

Chapter 4 introduced the random forest technique for modelling these vegetation patterns based on abiotical predictive variables, and Chapter 5 assessed the possibility to determine the most important environmental variables within this distribution modelling context. This chapter further builds on previously presented

---

The content of this chapter are published as J. Peters, B. De Baets, R. Samson and N. E. C. Verhoest. *Modelling groundwater-dependent vegetation patterns using ensemble learning.* Hydrology and Earth System Sciences, 12:603–613, 2008.

results, but focusses on model evaluation.

The two research questions under investigation in this chapter are:
1. Is there a spatial trend in the random forest distribution modelling results?
2. Does a random forest distribution model, constructed on a given wetland, perform satisfactorily when tested on a similar but distant wetland?

Therefore, a spatially explicit evaluation of the random forest distribution model predictions is made, followed by an assessment of the possibility to apply the random forest distribution model to a spatially distinct but similar ecosystem in independent model testing.

6.2 Ecohydrology of the Doode Bemde

During the summer of 1993 and the spring of 1994, plant species occurrences were mapped in the study area (Chapter 3). The total area of 21.08 ha was subdivided in 519 regular and adjacent 20 m by 20 m grid cells. Mapping was restricted to a selection of 56 plant species of which 45 were phreatophytes and 11 were differential species for several vegetation types at the Doode Bemde. Based on these species cover data, De Becker et al. [81] applied TWINSPAN [83] in order to define vegetation types. Seven different types were distinguished, and their spatial distribution can be seen in Figs. 4.7 and 6.1(a). All vegetation types are herbaceous, except for *Carici elongetae – Alnetum glutinosae* where a tree layer of Common Alder is present.

The similarity in species composition between grid cells was compared using the Jaccard index of similarity

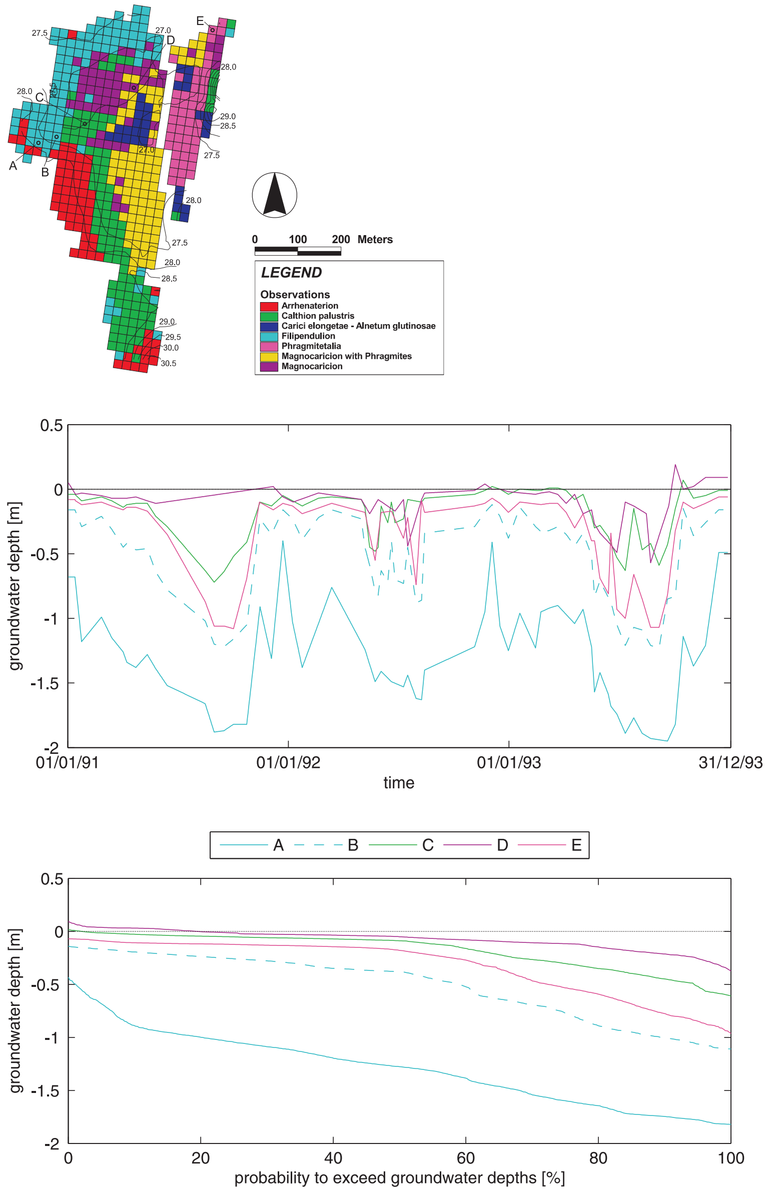
$$JS = \frac{c}{(a+b+c)} \tag{6.1}$$

where c is the number of species shared by both cells, and a and b are the numbers of species unique to each of the cells [262]. The Jaccard similarity of two grid cells expresses their ecological resemblance concerning species composition, and ranges between 0 (when both cells have unique species) and 1 (when both cells have equal species composition). Averaged JS values are given in Table 6.1 for the seven different vegetation types. The values of the diagonal elements in Table 6.1 are a measure of similarity between grid cells of the same vegetation type. Based on these values, patches of *Phragmitetalia*, *Magnocaricion* with *Phragmites* and *Magnocaricion* could be concluded to be more homogeneous in species composition compared to the other vegetation types which had lower values. Between the different vegetation types, marked differences in similarity could be observed. *Magnocaricion* with *Phragmites* had high similarities with *Phragmitetalia* and *Magnocaricion*. Between the other vegetation types, similarities were generally lower, but nevertheless differences could be observed. *Arrhenatherion elatioris* for example, had twice as much species in common with *Filipendulion* than with *Magnocaricion*.

TABLE 6.1 – Jaccard index of similarity between the vegetation types in the Doode Bemde. Vegetation types are abbreviated accordingly the List of Abbreviations and Acronyms.

	Ar	*Cp*	*Ce*	*Fi*	*Ph*	*MP*	*Ma*
Ar	0.40						
Cp	0.18	0.37					
Ce	0.11	0.17	0.46				
Fi	0.24	0.21	0.20	0.39			
Ph	0.09	0.19	0.35	0.22	0.55		
MP	0.10	0.19	0.30	0.23	0.44	0.51	
Ma	0.11	0.24	0.30	0.33	0.38	0.42	0.54

A groundwater monitoring network consisting of 25 piezometers was installed in 1989. Groundwater depths were measured every fortnight during the period 1/1/1991 – 31/12/1993. Time series of linearly interpolated groundwater depths measured at several piezometers (A–E, locations can be seen in Fig. 3.2) along a topographical transect are plotted in Fig. 6.1(b). A yearly pattern of larger depths in summer and more shallow groundwater in winter was observed at all piezometers. Based on these time series, hydrological duration lines expressing the probability [%] that a groundwater depth is exceeded were calculated (Fig. 6.1(c)). Groundwater depths corresponding to a probability of exceedance of 50% are yearly median groundwater depths. They differed considerably along the transect (Fig. 6.1(c)). At the levee near the river a median value of 1.27 m was measured (piezometer A), which decreased gradually moving further down toward the depression (piezometer B→C→D), with a minimal yearly median groundwater depth of 0.05 m measured at piezometer D in the center of the depression. Fig. 6.1(c) also shows different periods of superficial groundwater depths (< 0.3 m) in all piezometers, ranging from 75% of the year in piezometer C to 35% of the year in piezometers B and D. Groundwater depths measured in piezometer A were never smaller than 0.3 m. Additional to the monitoring of groundwater dynamics, all 25 piezometers were sampled on several groundwater quality variables during a sampling campaign in September 1993 with respect to pH, Cl^-, Ca^{2+}, Fe_{tot}, K^+, Mg^{2+}, NO_3^-–N, NH_4^+–N, $H_2PO_4^-$ and SO_4^{2-}. All values are in mg L^{-1} except for pH [-]. A soil type map was made based on 59 drillings to a depth of 1 m, evenly distributed over the study area. Management regime was assessed for each grid cell separately. Four different regimes could be distinguished (see Subsection 5.2.1).



(a)

(b)

(c)

FIGURE 6.1 – Vegetation distribution at Doode Bemde with piezometers (○) A–E along a topographical gradient (a) (see also Fig. 3.2). Time series of the groundwater depth, as monitored by piezometers A–E (b). Hydrological duration lines expressing the probability that measured groundwater depths are exceeded. The line colours correspond to the vegetation types wherein these piezometers were installed (c).

6.2.1 Data set

Groundwater depth measurements were used to calculate the average groundwater depth (AGD) below surface [m]. Values of this variable, together with the groundwater quality variables, were assigned to each grid cell by spatial interpolation of measurement data over the entire area using block kriging ([78, 263]).

The spatially explicit variables were structured into a data set. The data set contains $N = 519$ measurement vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ consisting of the values of $p = 13$ variables describing the abiotic environment:

- Groundwater quantity: average groundwater depth (continuous variable);
- Groundwater quality: pH, Cl^-, Ca^{2+}, Fe_{tot}, K^+, Mg^{2+}, NO_3^-–N, NH_4^+–N, $H_2PO_4^-$ and SO_4^{2-}. All these variables are continuous;
- Soil: soil type (silt/peat, categorical);
- Management: yearly mowing, cyclic mowing, transition from yearly to cyclic mowing, no management (categorical).

Seven different vegetation types $c_1, \ldots, c_7$ are considered. To each measurement vector $\mathbf{x}_i$ a unique vegetation type $l_i \in \{c_1, \ldots, c_7\}$ is assigned. The data set will be denoted as:

$$L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\} . \tag{6.2}$$

6.2.2 Independent evaluation data set

A spatially independent ecohydrological data set L_{ev} was constructed for a similar valley ecosystem, ‘Snoekengracht’. The Snoekengracht is an alluvial floodplain of the river Velp, situated approximately 15 km from the Doode Bemde. The climatic setting of both nature reserves is very much alike, and local environmental conditions and floral composition are very similar [78] (see Chapter 3). The monitoring scheme was largely the same as in the Doode Bemde ([72] and Chapter 3), and a grid-based (with a grid size of 10 m by 10 m) data set consisting of $M = 501$ elements was constructed, which will be denoted as:

$$L_{\mathrm{ev}} = \{(\mathbf{x}_{\mathrm{ev},1}, l_{\mathrm{ev},1}), \ldots, (\mathbf{y}_{\mathrm{ev},M}, l_{\mathrm{ev},M})\} , \tag{6.3}$$

where $l_{\mathrm{ev},i}$ is the vegetation type assigned to measurement vector $\mathbf{y}_{\mathrm{ev},i}$. Most vegetation types coincide with those found at Doode Bemde, except for *Magnocaricion* which was not found at Snoekengracht (see Table 3.2). The reason why not all 696 grid cells described for Snoekengracht were included in the data set L_{ev} came from the selection criteria used: only grid cells with a vegetation type that is present at Doode Bemde were included (grid cells with *Alno – Padion* excluded), and the management regimes present in the measurement vectors of L_{ev} should be one of the four management regimes applied at Doode Bemde (grid cells with transitional management from yearly mowing to no management and from cyclic mowing to no management were excluded). It could be argued to do the same for the continuous predictive variables, only to include values within the variable ranges found at Doode Bemde. This was not done here to guarantee a high degree of independence between training and test data set.

6.3 Modelling vegetation distributions

6.3.1 Model construction and results

First the data set L was randomly split into 3 data subsets for 3-fold cross-validation (following Algorithm 3 in Chapter 4, with $k = 3$). A random forest distribution model was constructed. User-defined parameters m, the number of randomly selected predictive variables to split the nodes, and k, the number of trees within the random forest, were optimized using the oob error, and suitable parameter values were $m = 3$ and $k = 1000$. The results include an ensemble of k (1000) predictions, one made by each classifier, which were aggregated based on majority votes into a final classification. A confusion matrix summarizing the final classification is given in Table 6.2, and results are shown in Fig. 6.2(a).

TABLE 6.2 – Confusion matrix of the classification made by the random forest distribution model. Predicted vegetation types are compared with the observations at the Doode Bemde.

		Observed						
		Ar	*Cp*	*Ce*	*Fi*	*Ma*	*MP*	*Ph*
Predicted	*Ar*	55	4	0	4	0	0	0
	Cp	6	89	0	7	4	5	0
	Ce	0	1	19	0	4	4	1
	Fi	9	2	0	82	7	0	1
	Ma	0	6	1	4	37	12	2
	MP	0	2	3	1	9	68	4
	Ph	0	2	7	1	2	4	45

6.3.2 Model evaluation

6.3.2.1 Classification accuracy

Out of the 519 grid cells of Doode Bemde included in the study, the model classified 395 (76.1%) correctly, and 124 (23.9%) incorrectly (Table 6.2). A κ [220] value of 0.716 was calculated, indicating a substantial agreement between observations and predictions. A threshold-independent evaluation using receiver operating

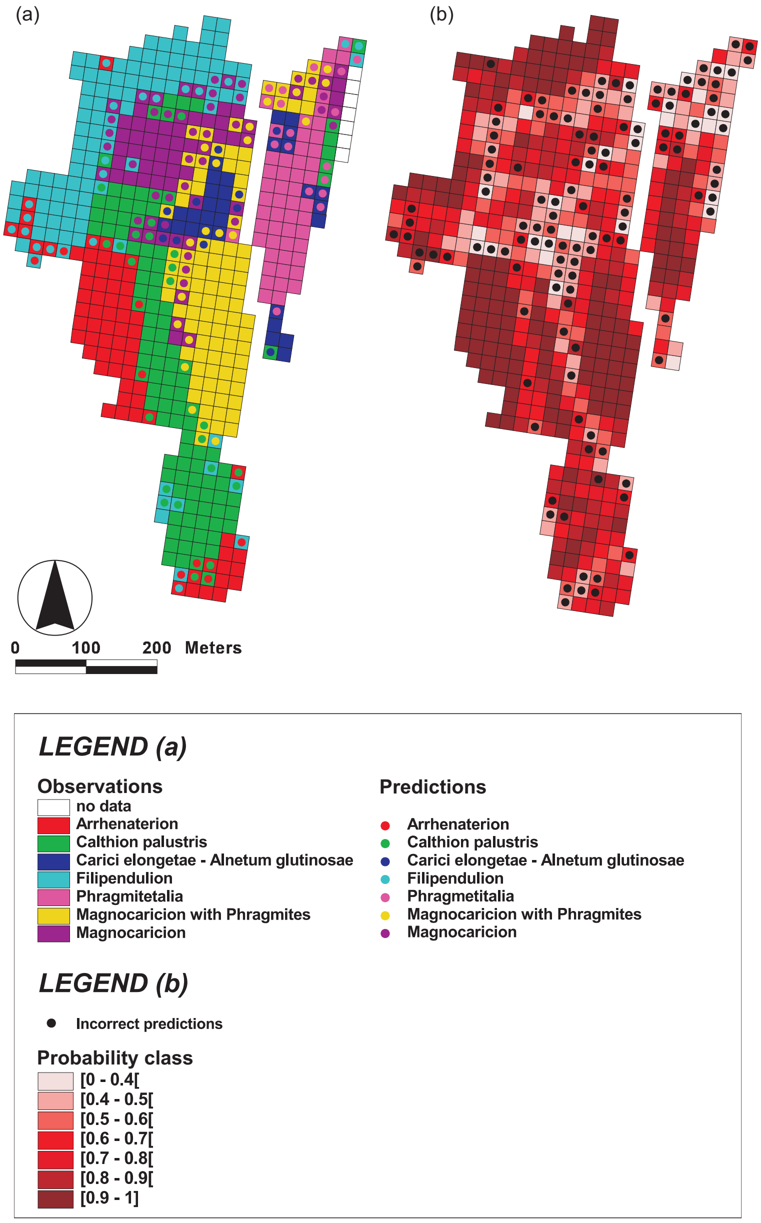


FIGURE 6.2 – Observed vegetation types overlaid by the classification made by the random forest distribution model (a). Modelled probabilities ($P(c)_{\max}$) on which the classification is based (b).

characteristic (ROC) curves was performed ([210] and Chapter 4). Recall that the area under the ROC curve (AUC) is a scalar value between 0 and 1 representing the classifier performance [227]. For multi-class ROC graphs, which should be applied here since 7 vegetation types are considered, a methodology described by Fawcett [227] is used. For each class a different ROC curve was produced, with ROC curve j plotting the classification performance using vegetation class c_j as positive and all other classes as negative. For each ROC curve, the AUC was calculated and averaged over the different classes using class weights based on class prevalences in the test data [264]:

$$\mathrm{AUC}_{\mathrm{total}} = \sum_{j=1}^{7} \mathrm{AUC}(c_j) \cdot w(c_j) \tag{6.4}$$

where AUC(c_j) is the area under the class reference ROC curve for c_j, and $w(c_j)$ a weighing factor. Weighing factors are obtained from Table 3.2. Fig. 6.3 visualizes the ROC curves for each vegetation type. The $\mathrm{AUC}_{\mathrm{total}}$ value equaled 0.96 and the random forest distribution model was concluded to perform well.

6.3.2.2 Spatially explicit evaluation

For each grid cell, the ensemble of k (=1000) classification results is aggregated by calculating probabilities of occurrence $P(c_j)$ for all j vegetation types of which the vegetation type with the highest $P(c_j)$ value ($P(c)_{\mathrm{max}}$) is the predicted one. As seen in Fig. 6.4, this decision rule led to an increasing number of correct classifications with increasing $P(c)_{\mathrm{max}}$ values. Indeed, 252 elements were correctly classified with a probability higher than 0.7, whereas only 2 elements were correctly classified with a probability lower than 0.3. 50% of the correctly classified elements were based on probabilities > 0.78. The incorrect classifications show a maximum in the [0.4,0.5[interval, with 1 element incorrectly classified with a probability lower than 0.3, and 28 elements incorrectly classified with probabilities higher than 0.7. 50% of the incorrectly classified elements were based on probabilities > 0.55.

Fig. 6.2(b) shows the spatial distribution of $P(c)_{\mathrm{max}}$ values at the study site in graduated colours. Correctly classified grid cells with high $P(c)_{\mathrm{max}}$ values were situated within the central areas of homogeneous vegetation clusters, and $P(c)_{\mathrm{max}}$ values tended to decrease toward the boundaries of these areas (Fig. 6.2(a)). Incorrectly classified grid cell are mainly found where two adjacent vegetation types meet, and were based on low $P(c)_{\mathrm{max}}$ values at the central depression and the north-eastern side of the study site. The vegetation types found in these areas are *Carici elongetae - Alnetum glutinosae*, *Phragmitetalia*, *Magnocaricion* with *Phragmites* and *Magnocaricion*. A Jaccard similarity matrix was constructed for the boundary grid cells only (Table 6.3). The *JS* values in Table 6.3 express averaged resemblances in species composition of each boundary grid cell with its

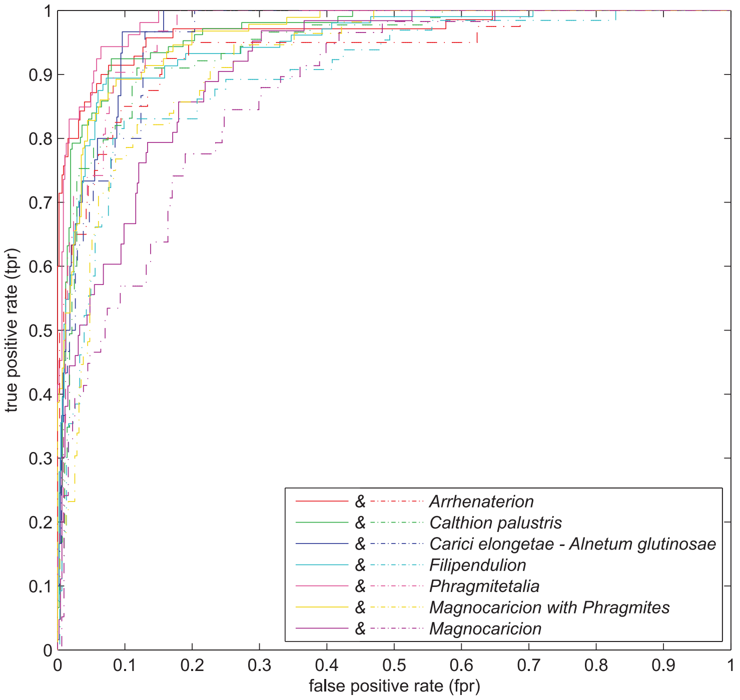


FIGURE 6.3 – Receiver operating characteristic (ROC) curves visualizing the classification performances of the 3-fold cross-validated random forest distribution model for the 7 vegetation types (full curves). The $\text{AUC}_{\text{total}}$ equals 0.96. Model performances for boundary cells only are summarized by the dashed ROC curves, yielding an $\text{AUC}_{\text{total}}$ value of 0.92.

neighbouring grid cells (maximal 8 neighbouring grid cells). Boundary grid cells of *Phragmitetalia*, *Magnocaricion* with *Phragmites* and *Magnocaricion* could be concluded to share a large proportion of their species with *JS* values higher than 0.5. This is reflected in the modelling results, $P(c)_{\max}$ values for these grid cells were generally low because comparable numbers of the $k = 1000$ classifiers classify these grid cells as *Phragmitetalia*, *Magnocaricion* with *Phragmites* and *Magnocaricion*. Another conclusion should be drawn for isolated grid cells and small isolated vegetation clusters surrounded by another vegetation type (e.g. as occurs along the western border of the study area, see Fig. 6.2a). These grid cells were frequently incorrectly classified with high $P(c)_{\max}$ values, and are a weak point of the random forest distribution model. The worse performance of the model on boundary grid cells could also be seen in Fig. 6.3, where ROC curves of classifica-

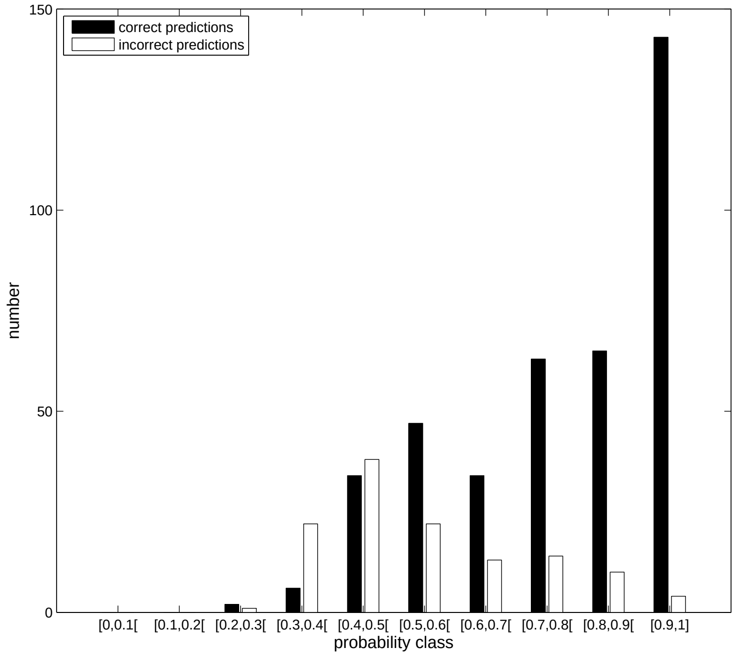


FIGURE 6.4 – Probability distribution of correct and incorrect classified grid cells of the Doode Bemde ($N = 519$).

tion results computed for boundary grid cells only were lower than those computed for the entire data set. The corresponding $\mathrm{AUC}_{\mathrm{total}}$ value for model performances in boundary areas equaled 0.92, while being 0.96 for the entire study area.

6.3.2.3 Performance on independent test data

The use of independent test data allows to assess the model generalization abilities. Edwards et al. [151] pointed out that cross-validated model accuracies are frequently different from accuracies assessed with truly independent data. It is easy to conclude that the random forest vegetation distribution model, which was trained on the data set L did not classify data set L_{ev} satisfactory. From the 501 elements included in L_{ev}, only 99 elements were classified correctly (19.8%). Two causes can attribute to this low level of model accuracy. A first cause can be best explained by the niche concept ([50], see Section 1.4). The fundamental niche of a plant species, and by extension a vegetation type, is defined as an n-dimensional

TABLE 6.3 – Jaccard index of similarity for boundary grid cells between two vegetation types at the Doode Bemde. Non-adjacent vegetation types are indicated by a dash.

	Ar	*Cp*	*Ce*	*Fi*	*Ph*	*MP*	*Ma*
Ar	0.59						
Cp	0.38	0.60					
Ce	–	0.45	0.66				
Fi	0.34	0.21	–	0.54			
Ph	–	0.18	0.52	0.27	0.67		
MP	–	0.30	0.36	0.19	0.57	0.65	
Ma	–	0.34	0.39	0.57	0.59	0.53	0.66

hypervolume [50] in which every point corresponds to a state of the environment which would permit the species to exist and reproduce. Due to interspecific interactions species generally occupy only an elementary part of this volume, the realized niche. The niches realized by each of the vegetation types found at the Doode Bemde differ from those realised by the same vegetation types at Snoekengracht and similar results were observed for all vegetation types. The example of *Calthion palustris* is given in Fig. 6.5. Since 13 environmental variables are used in this study, a principle component analysis (PCA) was performed to reduce dimensions and make results visible. Fig. 6.5 graphs the component scores of grid cells where *Calthion palustris* was observed on the first two principle component axes (cumulatively explaining 70% of variance). Although partly intersecting, two different realized niches can be distinguished. Obviously, a random forest distibution model that is trained on the vegetation distributions at the Doode Bemde and which uses explicit environmental thresholds to compute a classification, cannot perform well on such an independent test data set of an apparantely similar ecosystem.

A second cause of the low accuracy level of the independent modelling results lies in the model evaluation itself. Vegetation types were determined by species clustering by means of the TWINSPAN algorithm [83] for both sites independently. As can be seen in Table 6.1, grid cells of the same vegetation type do differ within the Doode Bemde (otherwise the diagonal elements of Table 6.1 would equal 1). This difference is even more pronounced for grid cells of the same vegetation type located in the two different study areas, Doode Bemde and Snoekengracht. Jaccard similarity values of 0.18, 0.35, 0.20, 0.19, 0.26, 0.25 and – (no value) were calculated for *Arrhenatherion elatioris*, *Calthion palustris*, *Carici elongatae – Alnetum glutinosae*, *Filipendulion*, *Phragmitetalia*, *Magnocaricion with Phragmites* and *Magnocaricion* (not found at Snoekengracht), respectively. These differences cannot be accounted for during the supervised model training (performed exclusively on Doode Bemde) but deteriorate the model performances on an independent data set.

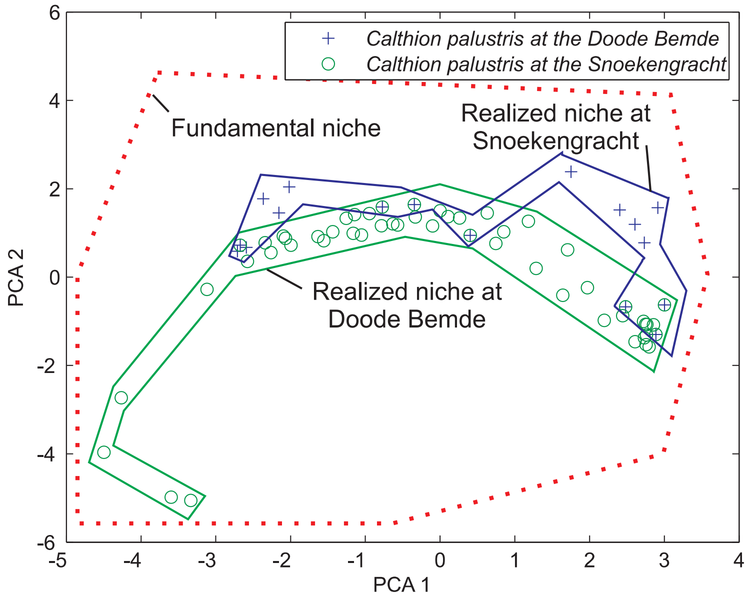


FIGURE 6.5 – Conceptual representation of realised niches of *Calthion palustris* at the Doode Beemde and Snoekengracht. The fundamental niche of *Calthion palustris* ranges over all environmental states which would permit to *Calthion palustris* to exist indefinitely [50].

6.4 Conclusions

Vegetation patterns arise from the interplay between intraspecific and interspecific biotic interactions and from different abiotic constraints and interacting driving forces and distributions [49]. In this chapter, a vegetation distribution model was constructed based on spatially distributed environmental variables which were linked with the occurrence of a certain vegetation type. Biotic interactions were only included indirectly, i.e. their effect was included through the observed vegetation distribution pattern, not directly as independent variables underlaying the vegetation distribution. Following conclusions could be drawn:

1. As far as classification accuracy of the random forest is concerned, results were satisfactory ($AUC_{total} = 0.96$).

2. Model errors were located in boundary areas ($AUC_{boundary\,area} = 0.92$) between adjacent vegetation types. A proportion of these errors could be attributed to high similarities between neighbouring grid cells. These incor-

rect predictions were generally based on low probabilities of occurrence of several similar vegetation types.

3. The random forest distribution model could not be applied beyond the local conditions upon which it was constructed, because realized niches of species/vegetation types do seldom coincide, even between apparently similar sites. This restricts the model's applicability. In order to make it operational on a larger scale many data would be needed, ranging over the entire ecological amplitude of the modelled vegetation types.

7

Assessing uncertainty propagation in the random forest distribution model

7.1 Introduction

Modelling of vegetation distributions across the landscape based on the relationship between the spatial distribution of environmental variables and vegetation is important for a range of management activities. Examples include management of threatened species and communities, risk assessment of non-native species in new environments, and the estimation of the magnitude of biological responses to environmental changes [265, 266]. In their attempt to summarize complex distributional patterns, however, distribution modelling results will inevitably contain some degree of uncertainty [266], and uncertainty assessment is gaining more and more attention in ecological modelling studies (e.g. [267–270]).

Uncertainty in vegetation distribution models originates from input data limitations, caused by spatial and temporal underrepresentation of observations to capture local variability, measurement errors on observations, systematic errors due to bias in the measurement equipment, missing of key environmental variables constraining the vegetation distribution, and subjective judgments, e.g. judgment on the type of environmental variables vegetation is sensible to, and their relative importance to classify vegetation types [266, 271]. Furthermore, distribution

---

The content of this chapter is submitted as J. Peters, N. E. C. Verhoest, R. Samson, M. Van Meirvenne, L. Cockx, Z. Vekerdy and B. De Baets. *Uncertainty propagation in vegetation distribution models based on ensemble learning.* Ecological Modelling, submitted, 2008.

modelling techniques introduce uncertainty by their disability to capture the entire complexity of ecological processes in relation to vegetation distributions. Distribution models are a simplified representation of the real world, and physical and biological processes are related frequently on empirical, statistical grounds. Finally, the model evaluation is susceptible to uncertainties.

Among this variety of sources of error and uncertainty, this chapter exclusively investigates two important sources of uncertainty propagating in vegetation distribution models: (i) the uncertainty associated with the spatial interpolation of environmental variables, and (ii) the uncertainty associated with species clustering into vegetation types. Other sources of error and uncertainty are not studied.

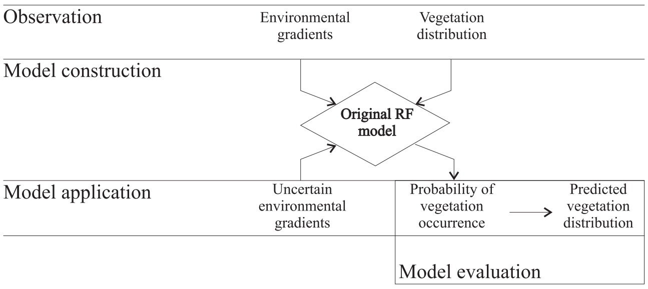
The two research questions under investigation in this chapter are:
1. *Does the use of an ensemble modelling technique allow for uncertainty assessment?*
2. *How does input uncertainty propagate through the random forest distribution model?*

Therefore, the potential of the random forest classifier ensemble for uncertainty assessment is investigated, followed by an uncertainty assessment associated with uncertain model input.

After the description of material and methods (Section 7.2), this chapter has a dichotomous structure imposed by the two different sources of uncertainty considered, and a conceptual difference in investigating their propagation through the distribution model. Firstly, attention was on the uncertainty associated with the spatial interpolation of environmental variables. A methodology (sequential Gaussian simulation) was applied to get an estimate of the local uncertainty associated with the spatial interpolation of environmental variables. A random forest distribution model was constructed on a training data set including the median simulated value for each environmental variable for each grid cell, together with its vegetation type, as originally determined. By calling this distribution model the *original random forest model*, the assessment strategy for uncertainty propagation due to uncertainty in environmental gradients is represented graphically in Fig. 7.1(a).

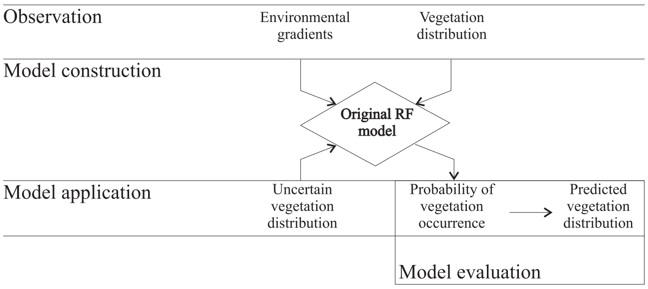
Secondly, for the propagation assessment of uncertainty associated with species clustering into vegetation types, a similar strategy could be followed (Fig. 7.1(b1)), i.e. the application of an uncertain vegetation distribution to the original random forest distribution model. This was not performed, however, since the model performs classification based on environmental variables exclusively, which means that response labels (the vegetation types of test instances) are not taken into account. Therefore, the model performance could be determined to a large extend beforehand. Assume a random forest distribution model gaining

(a) Uncertainty associated with the spatial interpolation of environmental variables.



(b) Uncertainty associated with species clustering into vegetation types.

(b1) Uncertain vegetation distribution applied to the original RF model.



(b2) Uncertain vegetation distribution in model construction and application.

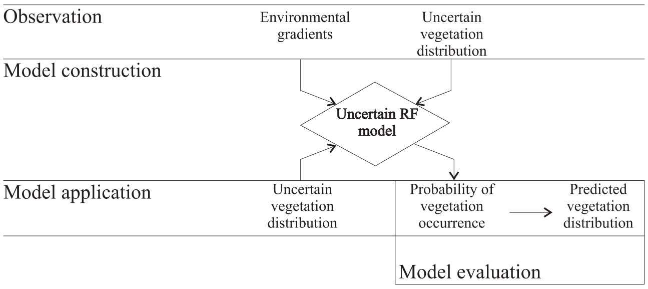

Observation
Environmental gradients
Uncertain vegetation distribution
Model construction
Uncertain RF model
Model application
Uncertain vegetation distribution
Probability of vegetation occurrence
Predicted vegetation distribution
Model evaluation

FIGURE 7.1 – Assessment strategy for uncertainty propagation due to uncertainty associated with the spatial interpolation of environmental variables (a), and associated with species clustering into vegetation types (b). The latter was investigated by strategy (b2), (b1) was not applied.

a perfect fit, than the application of uncertain test data would propagate linearly through to model, resulting in model performances that are directly proportional to the test data uncertainty. Such a trivial exercise would not gain the required insight on uncertainty propagation due to species clustering.

To meet that objective, another strategy was followed. Starting from the vegetation distribution which has been used throughout this dissertation (and which might include clustering errors as well), a known degree of uncertainty was introduced into the vegetation distribution by pseudo-randomizations. Then, a random forest distribution model was constructed on this uncertain response, and cross-validated against its independent and uncertain test data set (Fig. 7.1(b2)).

7.2 Material and methods

7.2.1 Study area and data set

Doode Bemde (see Chapter 3) was selected as the study area in this chapter. However, the environmental variables used in this chapter is a subset of the ones previously used. Chapter 5 indicated that random distribution models constructed on a reduced number of predictive variables did perform satisfactorily when ≥ 6 predictive variables were included. In this chapter, it was decided to use the seven most important predictive variables (in order to gain a satisfying goodness-of-fit of the reduced distribution model), based on the ranking determined by the 'variable importance measure' in Chapter 5 (Fig. 5.6). However, one alteration was made: minimal groundwater depth was replaced by average groundwater depth. This change may be justified by (i) the high correlation between both variables ($\tau = 0.715$, Table 5.1), (ii) the identification of average groundwater depth as the eighth most important predictive variable. The reason why it was decided to use average groundwater depth instead of minimal groundwater depth, arises from the frequent use of average variable values in ecohydrological distribution modelling, and in that sense, the introduction of uncertainty by averaging is an interesting research topic.

The observations used in this chapter were derived from a groundwater monitoring network consisting of 24 piezometers, of which 21 piezometers were located within the borders of the Doode Bemde, and 3 were installed on selected locations just outside the nature reserve. Groundwater depths [m] were measured every fortnight during the period 1/1/1991 – 31/12/1993. Furthermore, all 24 piezometers were sampled on several groundwater quality variables during two different sampling campaigns in 1993 with respect to pH [-], Cl^- [mg L^{-1}] and SO_4^{2-} [mg L^{-1}]. Topsoil samples were taken once at 59 locations, and the organic matter content of the samples was determined (Chapter 3). Management regime was assessed for each grid cell separately (Chapter 3), and four different regimes could be distinguished (Chapter 5–6). Plant species mapping (presence/absence) was done for each of the 519 grid cells, and restricted to a short list of species (Chapter 3 and part of Appendix A).

7.2.2 Variation partitioning in species data

Spatial autocorrelation is a very general property of ecological variables [217]. Spatial structures observed in ecological communities arise from two independent processes [217, 272]: (i) environmental variables that influence species distributions are usually spatially distributed, and (ii) ecological communities at any given locality are most often influenced by the community structure at surrounding localities, because of biotic processes such as growth, reproduction, mortality and migration. Variation partitioning [273–275] can be used to assess the importance of these two sources of spatial structure. Variation partitioning starts with coding of the spatial information using the principle coordinates of neighbour matrices (PCNM) approach, which is based on a principle coordinate analysis (PCoA, a.k.a. classical multidimensional scaling [276]) of a truncated matrix of geographic distances, and described in detail by Borcard and Legendre [274] and presented graphically in Fig. 7.2. Eigenvectors of the positive eigenvalues of the decomposed distance matrix are then used as spatial variables in a direct gradient analysis such as partial canonical ordination (e.g. redundancy analysis, RDA [194, 195] or canonical correspondence analysis, CCA [193]). Partial canonical analysis allows to partition the total variation in the species data into the following four parts [273]:

(a) The non-spatial environmental variation in the species data, which is the fraction of the species variation that can be explained by the environmental variables independently of any spatial structure;

(b) The spatial structuring in the species data that is shared by the environmental data;

(c) The spatial patterns in the data that are not shared by the environmental data included in the analysis;

(d) The fraction of species variation explained neither by spatial nor by environmental variables;

and can be represented graphically (Fig. 7.3).

7.2.3 Spatial interpolation using sequential Gaussian simulation

Point observations of environmental variables were spatially modelled using sequential Gaussian simulation (sGs, [277]), mainly because of its ability to model

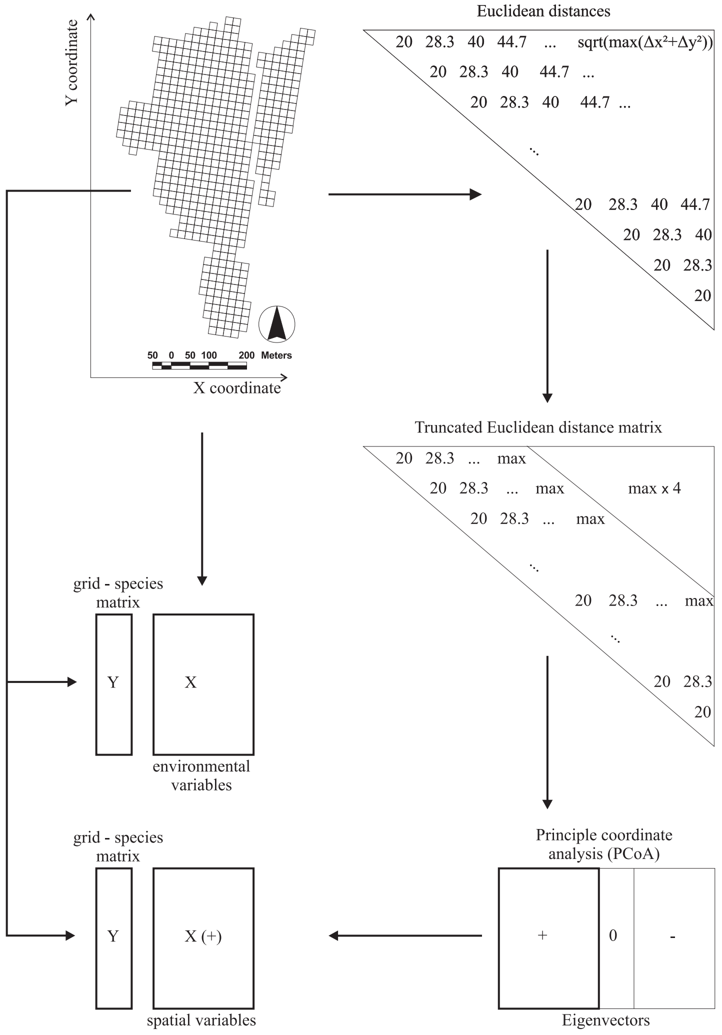


FIGURE 7.2 – The construction of spatial variables starts from spatial information (X and Y coordinates) which is used to calculate Euclidean distances. Principle coordinate analysis of the truncated distance matrix (with a given maximal distance (max)) results in a number of positive eigenvaulues which are used as spatial variables in a direct gradient analysis (adapted from [274]).

local uncertainty. Additionally, sGs preserves the characteristic roughness in the data, not producing a smoothed estimate but a reproduction of the real variability [278]. The sGs algorithm for the simulation of a single continuous random variable Z at N grid nodes $\mathbf{u}_j$ $(j = 1,\ldots,N)$ conditional to the observations of that variable $\{z(\mathbf{v}_\alpha), \alpha = 1,\ldots,n\}$ amounts to modelling the conditional cumu-

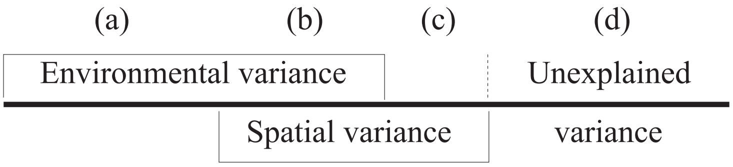


FIGURE 7.3 – Variation partitioning showing the 4 different fractions (adapted from [273]).

lative distribution function (ccdf) of that variable $F_{\mathbf{u}_j}(z|l) = P(Z(\mathbf{u}_j) \leq z|l)$. To ensure reproduction of the z-semivariogram model, each ccdf is made conditional to local information ($|l$) not only including the observations but also to values simulated at previously visited locations. The sGs algorithm is well described by Bourennane et al. ([279], modified in Algorithm 6) and Fagrout and Van Meirvenne [280] provided a flow-chart (adapted in Fig. 7.4). The sGs algorithm is available in the public domain [281].

The knowledge of the ccdf $F_{\mathbf{u}_j}(z|l)$ allows for local uncertainty assessment. If validation z-observations are available at N_V test locations $\{z(\mathbf{u}_j), j = 1, \ldots, N_V\}$, comparison of the median simulated value $F_{\mathbf{u}_j}^{-1}(0.5)$ and the observed validation value $z(\mathbf{u}_j)$ at the test locations allows for the examination of the bias and accuracy of the sGs algorithm is made. This examination is done by means of scatter diagrams of observed versus median simulated values at each test location, and by calculating error measurements, such as linear correlation coefficient (r), mean absolute error (MAE), and root mean square error (RMSE). Additionally, [282] developed a methodology to assess local model uncertainty visually. For a set of validation z-observations at N_V test locations $\mathbf{u}_j$ together with their corresponding, independently derived ccdfs $F_{\mathbf{u}_j}(z|l)$, $j = 1, \ldots, N_V$, the fraction of true values falling into the symmetric p-probability interval (PI) bounded by the $(1-p)/2$ and $(1+p)/2$ quantiles of their corresponding ccdf can be computed as:

$$\bar{\xi}(p) = \frac{1}{N_V} \sum_{j=1}^{N_V} \xi_j(p) \tag{7.1}$$

for any $p \in [0,1]$, with:

$$\xi_j(p) = \begin{cases} 1, & \text{if } F_{\mathbf{u}_j}^{-1}((1-p)/2) < z(\mathbf{u}_j) \leq F_{\mathbf{u}_j}^{-1}((1+p)/2), \\ 0, & \text{otherwise.} \end{cases} \tag{7.2}$$

The accuracy plot, which is a scatter diagram of the estimated ($\bar{\xi}(p)$) versus expected fractions (p), reflects the model accuracy: the model is accurate when the scatter points fall on or above the 1:1 line, and inaccurate when the points fall below the 1:1 line. In addition to model accuracy, one wants to know more about the model precision. Therefore, a precision plot has been proposed [282] in which, for

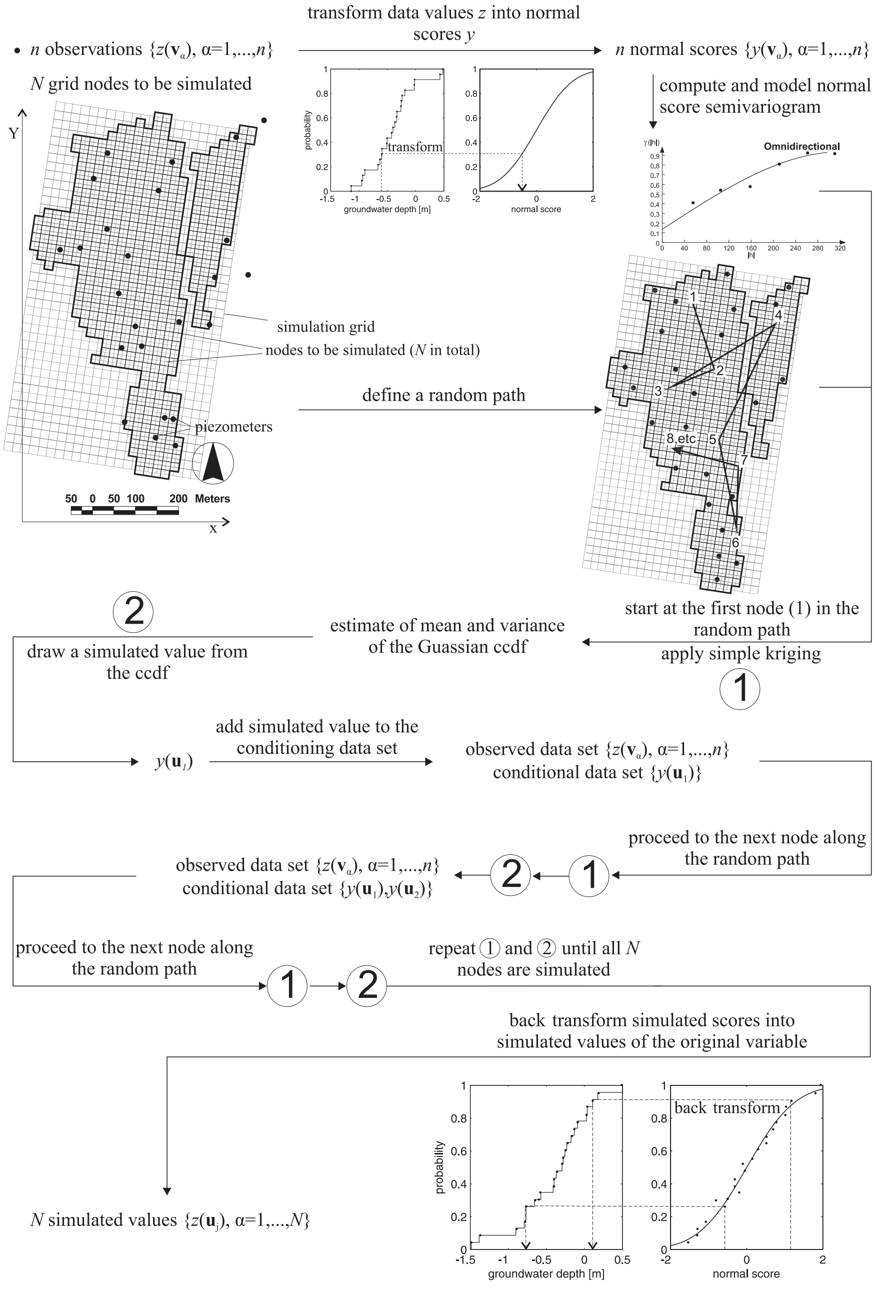

n observations {z(vα), α=1,...,n}
transform data values z into normal scores y
n normal scores {y(vα), α=1,...,n}
N grid nodes to be simulated
Y
x
simulation grid
nodes to be simulated (N in total)
piezometers
50 0 50 100 200 Meters
probability
transform
groundwater depth [m]
normal score
compute and model normal score semivariogram
γ(|h|)
Omnidirectional
|h|
define a random path
1
2
3
4
5
6
7
8,etc
start at the first node (1) in the random path
apply simple kriging
1
estimate of mean and variance of the Guassian ccdf
2
draw a simulated value from the ccdf
y(uj)
add simulated value to the conditioning data set
observed data set {z(vα), α=1,...,n}
conditional data set {y(u1)}
proceed to the next node along the random path
observed data set {z(vα), α=1,...,n}
conditional data set {y(u1),y(u2)}
2
1
proceed to the next node along the random path
1
2
repeat ① and ② until all N nodes are simulated
back transform simulated scores into simulated values of the original variable
N simulated values {z(uj), α=1,...,N}
probability
groundwater depth [m]
back transform
normal score

FIGURE 7.4 – Flowchart of sequential Gaussian simulations for groundwater depth at Doode Bemde (adapted from [280]). sGs are made for the nodes of a simulation grid, having equal size and orientation as the original grid, with simulation nodes located in the centre of the original grid cells.

Algorithm 6: The sequential Gaussian simulation (sGs) algorithm.

```
Data: set of observations $\{z(\mathbf{v}_\alpha), \alpha = 1, \ldots, n\}$ at locations $\mathbf{v}_\alpha$
      $(\alpha = 1, \ldots, n)$ of random variable $Z$
Result: simulation of a single continuous random variable $Z$ at $N$ grid nodes
define the number of realizations ($r$) as $k$;
for $r$ = 1 to $k$ do
    transform the observation data set $\{z(\mathbf{v}_\alpha), \alpha = 1, \ldots, n\}$ into normal
    scores $\{y(\mathbf{v}_\alpha), \alpha = 1, \ldots, n\}$ using the normal score transform:
        $y(\mathbf{v}_\alpha) = G^{-1}[\hat{F}_{\mathbf{v}_\alpha}(z)], \quad \alpha =, 1 \ldots, n$
    where $G^{-1}(\cdot)$ is the inverse Gaussian cumulative distribution function,
    and $\hat{F}$ is the sample cumulative distribution of $z$;
    compute and model the semivariogram of the normal scores ($\hat{\gamma}(h)$);
    define a random path along the nodes $\mathbf{u}_j$ $(j = 1, \ldots, N)$, visiting each
    node once;
    for $j$ = 1 to $N$ do
        (1) determine mean and variance of the Gaussian ccdf $G_{\mathbf{u}_j}(y|(c))$
        using simple kriging with the normal score semivariogram model
        $\hat{\gamma}(h)$. The conditional information $c$ consists of the normal score
        data $\{y(\mathbf{v}_\alpha), \alpha = 1, \ldots, n\}$ and values simulated at previously visited
        grid nodes $y^{(r)}(\mathbf{u}_a)$, with $a = 1, \ldots, j$ ;
        (2) draw a simulated value $y^{(r)}(\mathbf{u}_j)$ from that ccdf ;
        (3) add the simulated value to the conditioning data set ;
    end
    back transform the simulated normal scores $\left\{y^{(r)}(\mathbf{u}_j), \alpha = 1, \ldots, N\right\}$
    into simulated values of the original variable $\left\{z^{(r)}(\mathbf{u}_j), \alpha = 1, \ldots, N\right\}$
    by applying the inverse normal score transform to the simulated normal
    scores:
        $z^{(r)}(\mathbf{u}_j) = \hat{F}^{-1}[(G(y^{(r)}(\mathbf{u}_j))], \quad j = 1, \ldots, N$
end
```

a series of probabilities p, the average width of the PIs that include the observed values are plotted. The average width $\overline{W}(p)$ is computed as:

$$\overline{W}(p) = \frac{1}{N_V \overline{\xi}(p)} \sum_{j=i}^{N_V} \xi_j(p) \cdot [F_{\mathbf{u}_j}^{-1}((1+p)/2) - F_{\mathbf{u}_j}^{-1}((1-p)/2)] \tag{7.3}$$

and should be as small as possible for precise interpolations.

7.2.4 Species clustering

Cluster analysis of ecological data is an explicit way of identifying groups in data to find structures [192]. There are several clustering methods and a major distinction can be made between divisive and agglomerative methods. Divisive methods start with one group which is subsequently divided into smaller groups until a 'stopping rule' is satisfied. Agglomerative methods start with the individual objects which are combined into groups by collection of objects and groups into larger groups. In order to cluster species cover data into vegetation types, the TWINSPAN [83] program (divisive clustering) is frequently used in community ecology [192]. TWINSPAN produces a clustering of sites and species, by generating a two-way ordered table from a sites-by-species matrix. Within the two-way ordered table, the relative cluster similarity is given by a hierarchy of integer levels [283], and sites are clustered based on their species composition, and species are clustered into different vegetation types.

Additionally, a posterior analysis of the TWINSPAN site clustering results can be performed using the Jaccard index of similarity $JS = c/(a+b+c)$ where c is the number of species shared by both sites, and a and b are the numbers of species unique to each of the sites ([262] and Chapter 6). The Jaccard similarity of two sites expresses their ecological resemblance concerning species composition, and ranges between 0 (when both sites have unique species) and 1 (when both sites have equal species composition).

7.2.5 The random forest distribution model

The random forest model constructs an ensemble of k classification trees during model training. A unique class is assigned to a given data point by each of the k classification trees. The proportion of votes for a certain class $c_j \in C = \{c_1, \ldots, c_n\}$ over all k trees is interpreted as the probability of occurrence of that class:

$$P(c_j) = N_{c_j}/N_{\mathrm{tot}}\,, \tag{7.4}$$

with N_{c_j} the number of trees classifying the data point into class c_j, and N_{tot} $(= k)$ the total number of classification trees in the random forest. Thus, the random forest model output is a discrete probability distribution over all classes $c_j \in C$. The final classification is obtained by majority voting: the class with the highest probability of occurrence ($P(c)_{\max}$) is the predicted one. The uniformity of the discrete probability distribution allows to gain some information on model output uncertainty. Therefore, the Shannon entropy measure (H, [284, 285]), which has been applied in other ecological modelling studies (e.g. [286, 287]), can be used:

$$H = -\frac{1}{\log_2 n} \sum_{j=1}^{n} P(c_j) \log_2 P(c_j)\,, \tag{7.5}$$

with n the number of classes.

The value of H ranges between:

(i) 0 : when an identical class results from the classification of a given data point by every member of the random forest ensemble, i.e. the model output consists of probability values $P(c_j) = 1$, with $j \in \{1, \ldots, n\}$ and $P(c_k) = 0$, with $k = 1, \ldots, n$ and $k \neq j$; the $P(c)_{\max}$ value equals 1;

(ii) 1 : when the classification of a given data point results in any of the n different possible classes by equal numbers of members of the random forest ensemble, i.e. the model output consists of the following probability values $P(c_j) = 1/n$, with $j = 1, \ldots, n$; the $P(c)_{\max}$ value equals $1/n$;.

Within the context of vegetation distribution modelling, a value of H close to 0 indicates that, based on the environmental conditions of location i described in measurement vector $\mathbf{x}_i$, the random forest distribution model provides a strong expectation of a certain vegetation type. On the contrary, a value close to 1 indicates that, based on the environmental conditions, the random forest distribution model is not able to distinguish between the different vegetation types.

It should be stressed that the random forest distribution model generates an ensemble of classifiers. The discrete probability distribution over all classes resulting from classification by this ensemble suits for uncertainty assessment of the model output, an assessment that could not be made if a single classifier distribution model (or more generally, any distribution model computing a single response) was applied.

7.2.6 Evaluation of distribution modelling results

In this chapter, 3-fold cross-validation (see Chapter 4, Algorithm 3) was applied. Measures to evaluate the distribution modelling results are previously described. They include the oob error (see Chapter 4, Algorithm 2), which is defined as $(1-$accuracy of the classification of oob elements$)\times 100$ [%] and the test set error, which is defined as $(1-$accuracy of the classification of cross-validation test elements$)\times 100$ [%], where accuracy is the number of correctly classified instances divided by the total number of instances. Further, Cohen's κ test ([220] and Section 4.5) was used to evaluate differences between observations and predictions. The value of κ is negative if the agreement between observations and predictions is worse than expected by chance, and reaches 1 in case of perfect agreement.

A threshold-independent evaluation using receiver operating characteristic - (ROC) graphs was also performed [210] for visualizing classifier performance. For each ROC curve, the area under the curve (AUC) was calculated and averaged over the different classes using class weights based on class prevalences in the test data to obtain $\text{AUC}_{\text{total}}$ ([264] and Eq. (6.4)).

7.3 Uncertainty assessment related to uncertainty in environmental variables

As stated in the introduction, this chapter has a dichotomous structure. Under this section, an uncertainty assessment will be performed which is exclusively related to the uncertainty in environmental variables. The structure of this chapter is straightforward. The starting point are the field measurements (as made by the monitoring scheme described in Chapter 3). These measurements include environmental point measurements and area covering information on management and plant species distributions. The environmental point measurements are spatially interpolated, and the uncertainty associated with this interpolation is quantified. After model construction, an assessment is made of this source of uncertainty propagating through the model (as conceptualized in Fig. 7.1(a)).

7.3.1 From field observations to a spatially distributed data set

Field observations concerning the environmental site conditions were made at different ecosystem compartments: (i) groundwater (dynamics and quality), (ii) soil, and (iii) vegetation. Groundwater dynamics were described by a time series of groundwater depth measurements, while groundwater quality was described by means of concentration measurements of chemical groundwater compounds. Soil monitoring comprised the measurement of soil organic matter, and the direct anthropogenic impact on the vegetation compartment was assessed by identification of the different vegetation management regimes. The spatio-temporal density of field observations differed (Table 7.1), e.g. management regime was described for every grid cell of the study area ($N = 519$) on a single occasion, while groundwater depth observations were made 26 times each year (every fortnight) in 24 piezometers ($n = 24$) scattered over the area.

Based on these observations, seven different environmental variables were calculated, including average groundwater depth (AGD), amplitude of the groundwater depth (Ampli), pH, chloride concentration (Cl^{-}), sulphate concentration (SO_4^{2-}), soil organic matter content (SOM) and management regime (see Section 7.2.1). The former six variables are continuous, whereas the latter one is categorical with 4 possible management classes. Short summary statistics (mean, range, variance) of the environmental variables (Table 7.1) indicated marked hydrological differences within the study area, with average groundwater depths and groundwater amplitudes differences of more than 1.3 m between piezometers. Furthermore, groundwater quality as well as soil organic matter showed a high variability, and the study area could be concluded to comprise a high variability in environmental conditions. In addition to the environmental site observations, a species inventory, covering the entire study area, was made, i.e. for each of the

TABLE 7.1 – Spatio-temporal resolution of field observations made within different ecosystem compartments. Derived variables, abbreviations and summary statistics are included.

Ecosystem compartment	Measurement locations (n)	Measurement times per year (t)	Variable	Abbreviation	Unit	Summary statistics		
						mean	*range*	*variance*
groundwater depth	24	26	average groundwater depth	AGD	[m]	-0.45	[-1.35 -0.03]	0.12
groundwater depth	24	26	amplitude of groundwater depth	Ampli	[m]	1.06	[0.39 1.73]	0.11
groundwater quality	24	2	pH	pH	[-]	6.4	[5.7 6.7]	0.05
groundwater quality	24	2	chloride concentration	Cl^-	[mg L^{-1}]	24.1	[1.5 68.0]	223.1
groundwater quality	24	2	sulphate concentration	SO_4^{2-}	[mg L^{-1}]	53.5	[0.5 272.0]	3438.5
soil	59	1	soil organic matter content	SOM	[%]	20.7	[5.3 76.1]	290.1
vegetation	519	1	management regime	/	/	/	/	/

519 grid cells presence/absence records were made for all 56 plant species on the checklist (part of Appendix A). As stated in the introduction, all these field observations were assumed to be error free.

7.3.1.1 Variation partitioning in species cover data

To quantify the spatial component of ecological variation at the Doode Bemde, variation partitioning [273–275] was applied to 21 grid cells within the study area, i.e. those grid cells wherein field observations of groundwater dynamics and quality were made directly from a piezometer (from the 24 piezometers, 3 are located just outside the boundaries of the Doode Bemde). Three data sets (species, environmental and spatial, containing a grid-species matrix, environmental conditions and spatial information, respectively) were constructed (Fig. 7.2). The species data set consisted of inventory results of species occurrences (presence/absence) within each of the 21 grid cells. The environmental data set contained observations of AGD, Ampli, pH, Cl^- and SO_4^{2-} made from a piezometer within each of these 21 grid cells. Soil organic matter content of the nearest observation point, and management regime were added to the environmental data set. The spatial data set contained the 16 eigenvectors of the positive eigenvalues of the decomposed distance matrix. The species showed unimodal responses to the gradients in the study area (see Table 5.3, length of gradient > 2), and therefore the analysis was made using partial CCA. The whole variation of the species data set could be partitioned into the following parts: (i) non-spatially structured environmental variation, 20.8%; (ii) spatially structured environmental variation, 37.6%; (iii) spatial species variation that is not shared by the environmental data, 41.8%; and (iv) unexplained variation, 0.0%. Unexplained variance of 0.0% results from the high number of environmental and spatial variables where the species variation is explained upon.

The environmental variables explained 58.4% (37.6% + 20.8%) of the species variation, of which approximately two-thirds was explained by a similar spatial distribution of species and environmental variables, resulting partly from the same response of species and environmental variables to some common underlying causes. One-third of the explained species variation could be related to the environmental variables as such, and involved the local effect of these variables on plant species, without any spatial trend. 41.8% of the species variation was assessable by the spatial data set, and could not be related to any of the measured environmental variables. This means that unmeasured, but important environmental variables and processes, e.g. biotic processes of competition, predation and dispersal, were synthetically captured within the spatial data.

Variation partitioning indicated that the species distribution at the study area results from spatial distributions of both measured and unmeasured features. This result stresses the importance of an accurate spatial interpolation when species oc-

TABLE 7.2 – Summary of semivariogram models.

Variable	n	Model*	Nugget (C_0)	Sill ($C_0 + C_1$)	Range [m] (a)
AGD	24	sph	0.14	0.94	320
Ampli	24	exp	0.2	1	329
pH	24	sph	0.2	0.93	330
Cl^-	24	sph	0.1	0.95	348
SO_4^{2-}	24	exp	0.14	1.11	319
SOM	59	sph	0.17	1.08	297

*Models ($\gamma(0) = 0$)
Spherical (sph): $\gamma(h) = C_0 + C_1[3/2(|h|/a) - 1/2(|h|/a)^3]$ if $0 < |h| \leq a$
$\gamma(h) = C_0 + C_1$ if $|h| > a$
Exponential (exp): $\gamma(h) = C_0 + C_1[1 - \exp(-3|h|/a)]$ if $|h| > 0$

currence in relation with environmental conditions is under investigation. Furthermore, it indicates that there is uncertainty on the causality of the vegetation distribution, which makes the interpretation of the distribution modelling results harder. Finally, based on the variation partitioning result, the vegetation distribution model would probably benefit from the incorporation of spatial dependence [288], which was beyond the study objectives.

7.3.1.2 Uncertainty on spatial interpolation of environmental variables

The sGs algorithm was applied to the observation data set of each of the continuous environmental variables z (AGD, Ampli, pH, Cl^-, SO_4^{2-} and SOM) containing point measurements made at n locations $\mathbf{v}_\alpha$, $z(\mathbf{v}_\alpha), \alpha = 1, \ldots, n$. The normal score transformed z data were used to construct and model experimental omnidirectional semivariograms $\hat{\gamma}(h)$, with h the lag distance, using Variowin 2.2 software. Model parameters of the different semivariogram models are given in Table 7.2.

The simulations resulted in 500 back-transformed realizations for each variable for each of the 519 grid cells included in this study, based on which empirical non-parametric ccdfs were calculated (Fig. 7.5, example of groundwater depth). Median values ($\hat{F}(0.5)$) and conditional variances of these ccdfs were calculated. The conditional variance equaled 0 for grid cells where observations were made ($\hat{F}^{-1}(\cdot)$ = observed value). For other grid cells, values higher than 0 were calculated, and differences in values could be attributed to two main sources: (i) a spatial underrepresentation of nearby observations in the conditioning data set, and (ii) the presence of strong gradients in the conditioning data set, both resulting in highly variable estimates within the simulation algorithm. With respect to average groundwater depth, a spatial pattern could be observed in the conditional variance (Fig. 7.5). In the vicinity of the grid cells where observations were made, variance was generally low. Nevertheless, high variance values on the western levee with high average groundwater depths and in the central depression with superficial groundwater depths could be observed even in grid cells adjacent to the ones where observations were made, probably due to a lack of observation points within these areas. Similar variance patterns were found for the other continuous variables (not shown).

The lack of an independent validation data set forced the application of jackknifing to assess local uncertainty. Jackknife data sets (containing all but one observation) of the continuous variables AGD, Ampli, pH, Cl^-, $SO_4^{2-}$and SOM were applied to the sGs algorithm resulting in 500 realizations for each of the n_j jackknifed elements. The local uncertainty of the simulation results was investigated by means of scatter diagrams of observed versus median simulated values (Fig. 7.6). The error measurements indicated poor simulation results for most of the variables (AGD, Ampli, Cl^- and SO_4^{2-}), to moderate and good results for pH and SOM, respectively. Similar conclusions could be drawn from the accuracy plots. Scatter points were (partly) on or above the 1:1 line for pH and SOM, indicating accurate simulation results. The precision of the simulation results for these variables was also good. The width of the 0.5 probability interval was 0.22 units [-] and 13.84 [% org], for pH and SOM, respectively. The high local uncertainty of the simulation results of the other environmental variables could be attributed to the limited spatial coverage of observations.

For each grid cell i, the median value over all 500 realizations computed by the sGs simulation algorithm on the entire observation data set ($n = 24$ for all environmental variables, apart from SOM where $n = 59$) was taken for each continuous variable, and by adding management type which was identified for each of the grid cells separately, 519 measurement vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i7})$ constituted of the values of the seven spatially distributed environmental variables AGD, Ampli, pH, Cl^-, SO_4^{2-}, SOM and management type were constructed. To each measurement vector $\mathbf{x}_i$, a unique vegetation type $l_i \in \{c_1, \ldots, c_7\}$ was assigned to construct the data set $L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}$ with $N = 519$. The data set L will be used as a reference data set throughout this chapter.

Furthermore, for each grid cell i, 100 samples were drawn from the sGs simulation results for the continuous variables using Latin hypercube sampling [289, 290]. Latin hypercube sampling is a stratified random procedure that provides an efficient way of sampling variables from their distributions, by covering the full range of each variable [291]. Linked with the categorical variable management type and the observed vegetation type, 100 data sets were constructed ($LHS_1, \ldots, LHS_{100}$), in which the entire variable distributions are captured.

Where the Latin hypercube sampling resulted in data sets covering the entire probability ranges of the continuous variables, another, more systematic methodology was followed to construct data sets with varying degrees of deviation from the median simulation results. For each grid cell i, the median value and standard

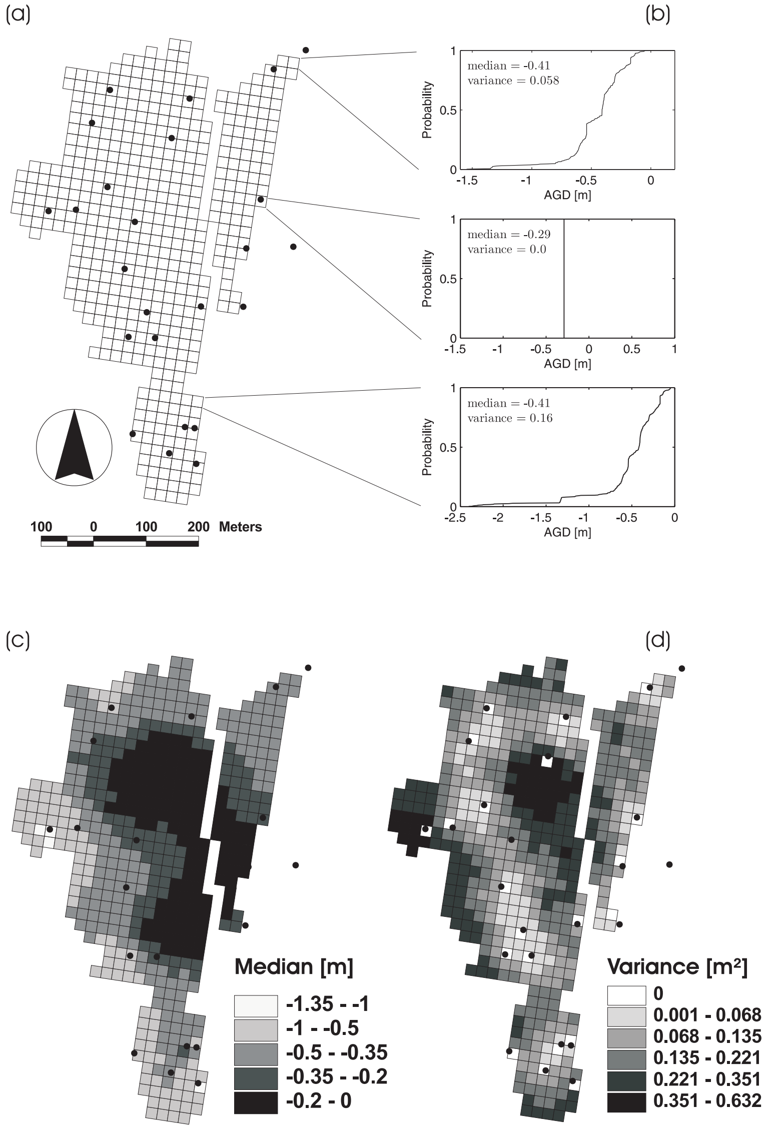


FIGURE 7.5 – Groundwater depths were monitored by piezometers (black dots, $n = 24$) scattered in and around the study area (a). Sequential Gaussian simulation using these observations resulted in 500 equiprobable groundwater depth realizations for each grid cell ($N = 519$). Empirical non-parametric conditional cumulative distribution functions (ccdfs) were computed from these realizations (b). Median (c) and variance (d) values were calculated based on the unique ccdf of each grid cell.

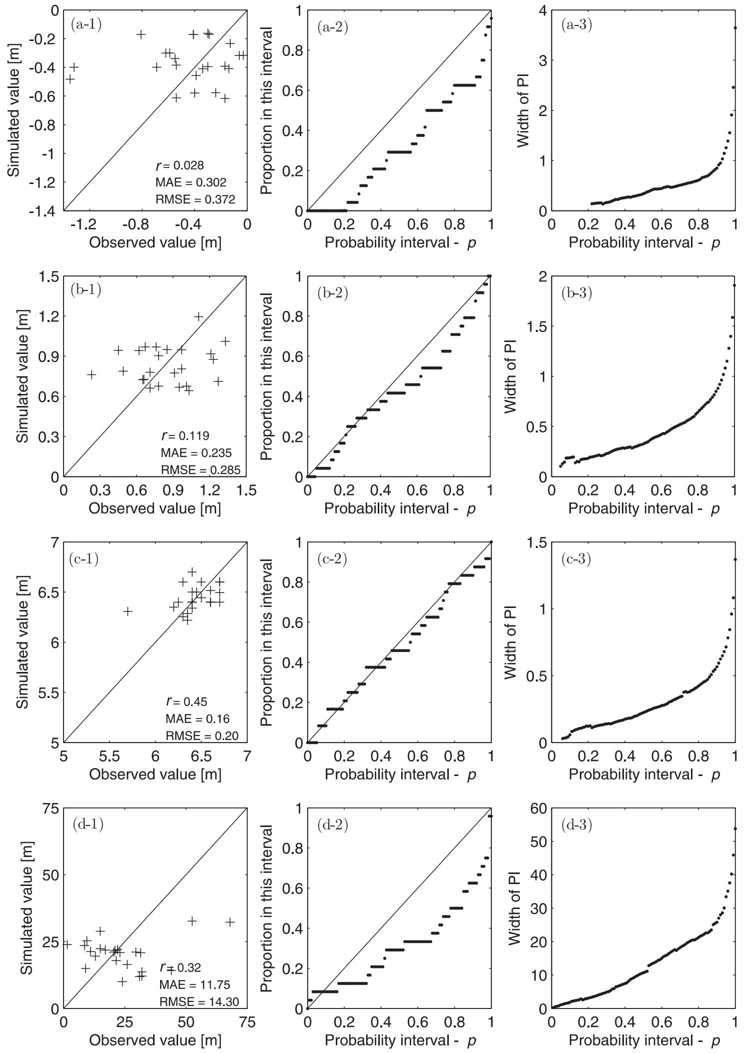


FIGURE 7.6 – Scatter diagram (1), accuracy plot (2) and precision plot (3) for the simulation results of the jackknifed elements of AGD (a), Ampli (b), pH (c), Cl^- (d), SO_4^{2-} (e), SOM (f).

deviations over all 500 realizations computed by the sGs simulation algorithm of the six continuous environmental variables were calculated. Data sets were created by adding a proportion a of the standard deviations to the median values, $L_{m+a\times stdev}$, and by subtracting a proportion a of the standard deviations from the median values, $L_{m-a\times stdev}$, where $a \in \{0, 0.01, 0.05, 0.1, 0.5, 1, 2\}$, resulting in 13 mutually exclusive data sets. When $a = 0$, the data sets $L_{m\pm 0}$ equal the reference

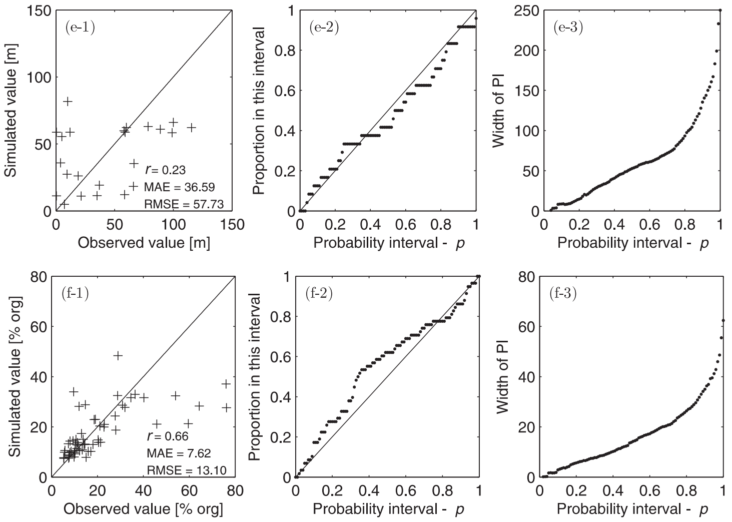


FIGURE 7.6 – *continued…*

data set *L*.

7.3.2 Model construction, calibration and evaluation

As indicated in the previous chapters (see Chapters 4–6), the number of trees (k) and the number of predictive variables used to split the nodes (m) are two user-defined parameters required to grow a random forest, which have to be calibrated to minimize model error. Parameter calibration can be done using the built-in out-of-bag error testing or using the test set error in cross-validation (see Chapter 4). In 3-fold cross-validation using different values of m, oob error and test set error were averaged over the three random forest models. Fig. 7.7 showed convergence of the random forest models constructed with different numbers of m ($m = 1$ (minimal value), $m = 3$ (optimal value), and $m = 7$ (maximal value)) when more trees are added (i.e. k increases). The values 1000 and 3 were used for the two user-defined parameters k and m, respectively.

Using these parameter values, a random forest distribution model was constructed (Algorithm 7) which made a classification of the 519 grid cells included in this study, of which 359 (69.17%) grid cells were classified correctly, and 160 (30.83%) grid cells incorrectly. A value of κ [220] of 0.633 was calculated, indicating a substantial agreement between observations and predictions. A threshold-independent evaluation using receiver operating characteristic (ROC) graphs was performed ([264] and Eq. (6.4)). For each class a different ROC curve was pro-

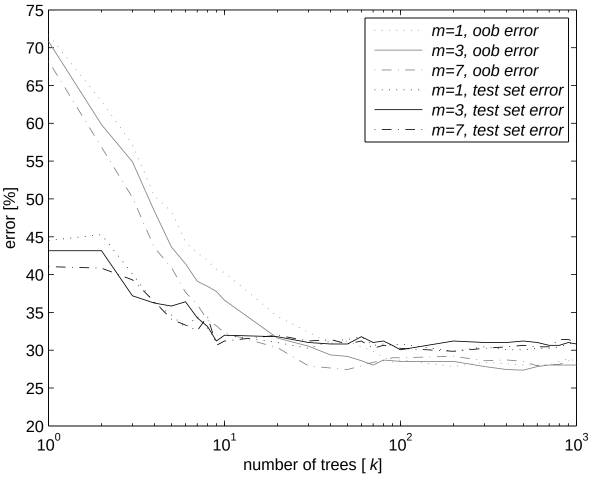


FIGURE 7.7 – Out-of-bag (oob) error and test set error converge when more trees are added to the random forest (when k increases). The numbers of variables (m) used to split the nodes are $m = 1$, $m = 3$ and $m = 7$. Average error values of the 3-fold cross-validated random forest model are plotted.

duced, with ROC curve j plotting the classification performance using vegetation class c_j as positive and all other classes as negative. For each ROC curve, the AUC was calculated and averaged over the different classes using class weights based on class prevalences in the dataset. The $\text{AUC}_{\text{total}}$ value equaled 0.943 and the random forest distribution model was concluded to perform well.

The random forest model output for each grid cell is a discrete probability distribution over the seven vegetation classes (see Eq. (7.4)). Looking into this probability distribution by means of Shannon's entropy measure H (Eq. (7.5)) allowed to gain some information on model output uncertainty. H values range between the maximal value 1 and minimal value 0. Other important H values are 0.356, 0.565, 0.712, 0.827 and 0.921, values obtained when the classification results include j dominant vegetation types with probabilities of occurrence $1/j$, where $j = 2, \ldots, 6$, respectively. When frequency counts were plotted against values of H computed for every grid cell in the study site (Fig. 7.8(a)), a decrease in frequency counts could be seen with increasing H values. This means that the random forest model output distribution was generally quite narrow, with a clear dominance of one, two or — to a lower extent — three different vegetation types.

Algorithm 7: Pseudo-code for random forest distribution model construction and testing using 3-fold cross–validation.

```
Data: $L$
Result: $P(c_j)$ values and test statistics $\kappa$, AUC$_{\text{total}}$ and $H$
partition the data set $L$ into 3 disjoint test data sets $T_1, T_2$ and $T_3$;
for $i = 1:3$ do
    use $S_i = L - T_i$ to construct random forest model RF$_i$;
    calculate the out-of-bag-error;
    save model;
    apply the saved random forest model RF$_i$ to test data set $T_i$;
    calculate the test set error;
    save $P(c_j) = N_{c_j}/N_{\text{tot}}$ (Eq. (7.4)) for all elements of $T_i$;
end
calculate test statistics $\kappa$, AUC$_{\text{total}}$ and $H$;
```

7.3.3 Uncertainty on spatial interpolation of environmental variables propagating to the modelling results

A random forest distribution model was constructed on the reference data set L, using calibrated parameters $k = 1000$ and $m = 3$ (Section 7.3.2). In 3-fold cross-validation, the model was then applied to two test data sets with quantified data uncertainty associated with the spatial interpolation of environmental variables: (1) the Latin hypercube test data sets, and (2) the deviation from median data sets.

7.3.3.1 Latin hypercube test data sets

The random forest distribution model applied to the Latin hypercube data sets ($LHS_1, \ldots, LHS_{100}$) using 3-fold cross-validation (Algorithm 8) resulted in probability of occurrence values for all seven vegetation types for each grid cell in the study area, and this for each of the 100 Latin hypercube test data sets. Modelling results were not accurate (Table 7.3); from the 100 test data sets containing 519 elements, on average only 229.7 elements (47.37%) were classified correctly (compared to 69.17% during model calibration), and a κ value of 0.367 and AUC$_{\text{total}}$ value of 0.828 were obtained (Table 7.3). Therefore it could be concluded that the model did not perform satisfactorily when the entire probability ranges of the continuous environmental variables are considered. A more detailed investigation of these modelling results was made by a grid-wise comparison of variances (Fig. 7.9). It was hypothesised that grid cells with low variances in simulation results for the continuous environmental variables (i.e. grid cells where observations are made, and simulated values equal the observed value, $\hat{F}^{-1}(\cdot)$ = observed value, as an extreme example) have a low variance in modelled probability of occurrence

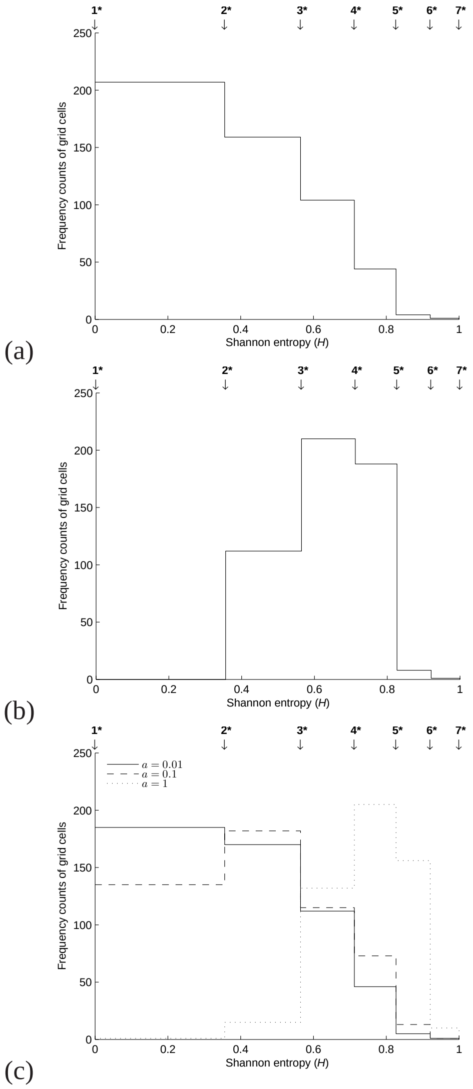


FIGURE 7.8 – Histogram of frequency counts of the Shannon entropy (H) values of the entire study site ($N = 519$) for the random forest distribution model cross-validated on L (a), and tested on the Latin hypercube samples (averaged) (b), and the deviation from median test data sets $L_{m+a\times stdev}$ with $a = 0.01$, $a = 0.1$, and $a = 1$ (c).
Legend: **j*** indicates the values of H obtained when a grid cell is classified as **j** vegetation types with equal probability of occurrence ($P(c_{\mathbf{j}}) = 1/\mathbf{j}$).

values. Therefore six scatter plots were constructed, plotting the variances in simulation results of the continuous variables against the variance in $P(c)_{\max}$ for the 100

Algorithm 8: Pseudo-code for model testing with Latin hypercube test data sets.

```
Data: $\mathrm{LHS}_k$, $k = 1, \dots, 100$
Result: $P_k(c_j)$ values and test statistics $\kappa_k$, $\mathrm{AUC}_{\mathrm{total},k}$ and $H_k$
for $k$ = 1:100 do
    use the partitioning of Algorithm 7 to partition the data set $LSH_k$ into 3 disjoint test data sets $T_{k,1}, T_{k,2}$ and $T_{k,3}$;
    for $i = 1:3$ do
        apply the saved random forest model $\mathrm{RF}_i$ to test data set $T_{k,i}$;
        calculate the test set error;
        save $P_k(c_j) = N_{k,c_j}/N_{\mathrm{tot}}$ (Eq. (7.4)) for all elements of $T_{k,i}$;
    end
    calculate test statistics $\kappa_k$, $\mathrm{AUC}_{\mathrm{total},k}$ and $H_k$;
end
```

Latin hypercube model testing runs for each grid cell. Four different groups were created within each plot based on model accuracy for each grid cell ($N = 519$): (1) a grid cell correctly classified in ≤ 25 on a total of 100 Latin hypercube test runs, (2) a grid cell classified correctly in ≤ 50 and > 25 model testing runs, (3) a grid cell classified correctly in ≤ 75 and > 50 model testing runs, and (4) a grid cell classified correctly in ≤ 100 and > 75 model testing runs. By applying Spearman's (r_s), correlations between variances were calculated from each of the four groups seperately. Significant positive correlations at the 0.05 significance level were found for grid cells that were classified correctly in $>$75 of the 100 test model runs. These include 19 grid cells where observations were made (located in the origin of the scatter plots). For the other groups, no significant correlations were found.

Calculation of the entropy of the random forest model output H for all grid cells i, averaged over all 100 Latin hypercube test runs, resulted in a histogram of frequency counts (Fig. 7.8(b)) showing a maximum between $H = 0.565$ and $H = 0.712$. Grid cells were mostly classified as three or four different vegetation types with similar probabilities of occurrence. Zero grid cells were classified with a H value < 0.356. In comparison with the histogram based on the cross-validated results of the reference random forest distribution model, a clear shift toward higher H values was observed, indicating that uncertainties on the spatial interpolation are propagated to the distribution modelling results.

TABLE 7.3 – Uncertainty on environmental variables propagating to the random forest distribution modelling results.

Data set		oob error [%]	test set error [%]	Cohen's κ	$\text{AUC}_{\text{total}}$	average H
L		28.03	30.83	0.633	0.943	0.420
LHS		27.55	52.63	0.367	0.828	0.661
$L_{m+a\times stdev}$	$a=0$	28.03	30.83	0.633	0.943	0.420
	$a=0.01$	28.03	29.67	0.646	0.941	0.435
	$a=0.05$	28.03	31.02	0.631	0.938	0.461
	$a=0.1$	28.03	33.53	0.600	0.931	0.491
	$a=0.5$	28.03	45.47	0.457	0.877	0.658
	$a=1$	28.03	53.76	0.363	0.829	0.768
	$a=2$	28.03	68.59	0.207	0.759	0.831
$L_{m-a\times stdev}$	$a=0$	28.03	30.83	0.633	0.943	0.420
	$a=0.01$	28.03	30.44	0.637	0.943	0.421
	$a=0.05$	28.03	31.41	0.624	0.935	0.433
	$a=0.1$	28.03	32.95	0.605	0.927	0.459
	$a=0.5$	28.03	53.18	0.347	0.860	0.615
	$a=1$	28.03	65.70	0.181	0.757	0.629
	$a=2$	28.03	70.91	0.123	0.678	0.623

7.3.3.2 Deviation from median test data sets

Another approach to assess uncertainty propagation to the modelling results made use of test data sets $L_{m\pm a\times stdev}$. In 3-fold cross-validation random forest distribution models were constructed on the reference data sets L_{12}, L_{13} and L_{23} and tested on according test data sets in which the values of the continuous variables differed in degree of deviation of the median simulated value, and for which the factor a is indicative (as it represents the proportion of the standard deviation that was added to and subtracted from the median values) (Algorithm 9).

Algorithm 9: Pseudo-code for model testing with deviation from median test data sets.

```
Data: $L_{m+a\times stdev}$, with
      $a \in \{-2,-1,-0.5,-0.1,-0.05,-0.01,0,0.01,0.05,0.1,0.5,1,2\}$
Result: $P_a(c_j)$ values and test statistics $\kappa_a$, $\text{AUC}_{\text{total},a}$ and $H_a$
for $a \in \{-2,-1,-0.5,-0.1,-0.05,-0.01,0,0.01,0.05,0.1,0.5,1,2\}$ do
    use the partitioning of Algorithm 7 to partition the data set $L_p^b$ into 3
    disjoint test data sets $T_{b,1}, T_{b,2}$ and $T_{b,3}$;
    for $i = 1:3$ do
        apply the saved random forest model $\text{RF}_i$ to test data set $T_{a,i}$;
        calculate the test set error;
        save $P_a(c_j) = N_{a,c_j}/N_{\text{tot}}$ (Eq. (7.4)) for all elements of $T_{a,i}$;
    end
    calculate test statistics $\kappa_a$, $\text{AUC}_{\text{total},a}$ and $H_a$;
end
```

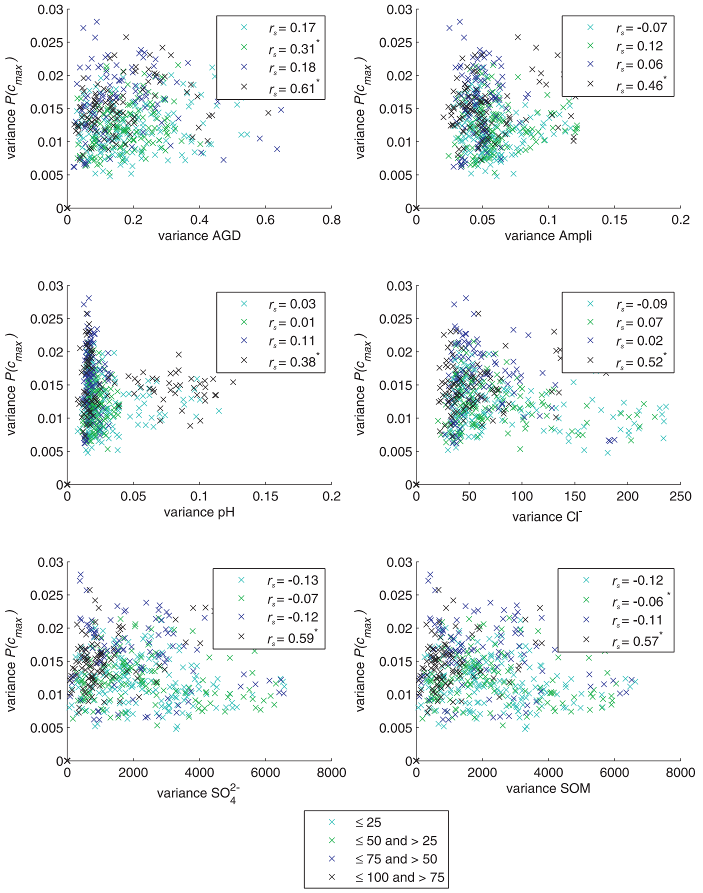


FIGURE 7.9 – Variable variance versus variance in modelled probability of occurrence of the predicted vegetation type ($P(c)_{\max}$) when the random forest distribution model is applied to 100 Latin hypercube test data sets. Different colours group the scatter points ($N = 519$) based on classification accuracy: a turquoise × is a grid cell classified correctly in ≤ 25 Latin hypercube model testing runs (out of 100), a green × is a grid cell classified correctly in > 25 and ≤ 50 model testing runs, a blue × is a grid cell classified correctly in > 50 and ≤ 75 model testing runs, and a black × is a grid cell classified correctly in > 75 and ≤ 100 model testing runs. Spearman's rank correlations (r_s) and significance at the 0.05 significance level (*) are indicated for each group separately.

Results indicated increasing test set errors when $|a|$ increased, and decreasing κ and AUC$_{\text{total}}$ values when $|a|$ increased (Table 7.3). However, the increase in test set error and decrease in κ and AUC$_{\text{total}}$ values was limited when $-0.1 \leq a \leq 0.1$, and much more pronounced when this threshold was exceeded. Based on this result it could be concluded that the random forest model performed well for test elements with low margins of deviation from training values. If a certain deviation threshold was exceeded, model errors increased drastically. The histogram with frequency counts of H values indicated a clear shift to higher values when test data deviated more from training data (Fig. 7.8(c)).

7.3.3.3 Implications for empirical distribution modelling

Distribution modelling on both Latin hypercube and deviation from median test data sets emphasized that environmental variables with low uncertainty are primordial for accurate distribution modelling. At the site scale, this amounts to increasing the monitoring density allowing accurate and precise spatial interpolation. The inclusion of stable environmental variables with limited spatial and temporal variability would lower the uncertainty on spatial interpolation as well, and could therefore be justified within an empirical distribution modelling context. The ability to explain vegetation patterns by such environmental variables, however, is questionable.

7.4 Uncertainty assessment related to uncertainty in vegetation distribution

As described in Chapter 3, vegetation types were determined on a spatial plant species inventory. Inevitably, species clustering introduces uncertainty in the vegetation distribution. This section assesses the propagation of such uncertainty to the distribution modelling results (as conceptualized in Fig. 7.1(b2)).

7.4.1 Uncertainty on species clustering

Based on the species cover data, TWINSPAN [83] was applied in order to define vegetation types (see Chapter 3 and Chapters 5–6). Seven different vegetation types were distinguished at the study site. A simplified representation of the TWINSPAN dendrogram is given (Fig. 7.10), and the spatial distribution of the seven different vegetation types can be seen in Fig. 3.2. A more detailed description of these vegetation types is included in Chapter 3, Table 3.2.

Uncertainty concerning the species clustering results from the many hard, arbitrary choices that had to be made. First of all, which clustering strategy is to be used: an agglomerative strategy or a divisive strategy? And if an agglomerative method is chosen, which (dis)similarity measure is to be used to base the clustering upon? Furthermore, what is the appropriate number of clusters? All these choices

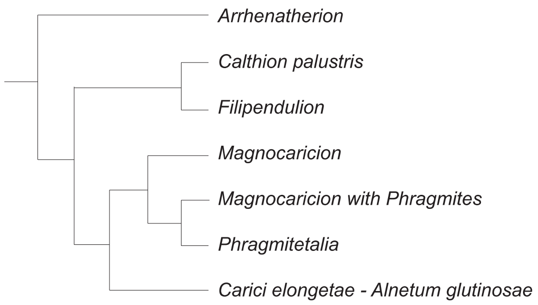


FIGURE 7.10 – Cluster dendrogram.

have to be made and influence the solution [292]. Additionally, the stability of the TWINSPAN solution is often of concern [292–294].

A posterior analysis of the TWINSPAN grid cell clustering was performed using the Jaccard index of similarity (JS). Averaged JS values are given in Table 6.1 for the seven different vegetation types, and as discussed in Chapter 6 marked differences in similarity between the different vegetation types is present.

Based on this analysis, 6 new data sets were constructed by pseudo-randomization of the response variable (vegetation type) of 1%, 5%, 10%, 20%, 50% and 100% of the N elements to assess the effect of uncertainty on the response variable. Pseudo-randomizations were based on the Jaccard similarity between grid cells of the seven different vegetation types (Table 6.1). This stategy reflects the likelihood of erroneous clustering of a grid cell based on its species composition. An *Arrhenatherion elatioris* grid cell for example, had on average approximately twice as much species in common with *Filipendulion* than with *Magnocaricion*; their respective JS values were 0.24 and 0.11. Therefore the likelihood is higher to classify the vegetation type of this grid cell as *Filipendulion* than as *Magnocaricion*. This difference was (linearly) taken into account during response pseudo-randomizations. The new data sets are referred to as L_p^b where subscript p refers to pseudo-randomization and superscript b to the percentage of pseudo-randomized elements.

7.4.2 Uncertainty on species clustering propagating to the modelling results

The data sets with pseudo-randomizations in the response variable (L_p^b) were used for model construction and testing (Algorithm 10). The reason why the calibrated model constructed on reference data set L was not used here, is that the random forest algorithm constructs its classifiers taking response variables into account (supervised learning), and hence uncertainty related to species clustering should be taken into account during model construction as well.

Algorithm 10: Pseudo-code for model testing with uncertainty on species clustering.

```
Data: $L_p^b$, with $b \in \{0,1,5,10,20,50,100\}$
Result: $P_b(c_j)$ values and test statistics $\kappa_b$, $\mathrm{AUC}_{\mathrm{total},b}$ and $H_b$
for $b \in \{1\%,5\%,10\%,20\%,50\%,100\%\}$ do
    use the partitioning of Algorithm 7 to partition the data set $L_p^b$ into 3 disjoint test data sets $T_{b,1}, T_{b,2}$ and $T_{b,3}$;
    for $i = 1:3$ do
        use $S_i = L_p^b - T_{b,i}$ to construct random forest model $\mathrm{RF}_i$;
        calculate the out-of-bag-error;
        apply $\mathrm{RF}_i$ to test data set $T_{b,i}$;
        calculate the test set error;
        save $P_b(c_j) = N_{b,c_j}/N_{\mathrm{tot}}$ (Eq. (7.4)) for all elements of $T_{b,i}$;
    end
    calculate test statistics $\kappa_b$, $\mathrm{AUC}_{\mathrm{total},b}$ and $H_b$;
end
```

TABLE 7.4 – Uncertainty on species clustering propagating to the random forest distribution modelling results.

Data set		oob error [%]	test set error [%]	Cohen's κ	$\mathrm{AUC}_{\mathrm{total}}$	average H
L		28.03	30.83	0.633	0.943	0.420
L_p^b	$b = 1\%$	29.48	29.29	0.652	0.942	0.417
	$b = 5\%$	30.83	32.56	0.613	0.919	0.456
	$b = 10\%$	37.96	37.76	0.551	0.860	0.523
	$b = 20\%$	49.04	49.71	0.413	0.791	0.626
	$b = 50\%$	76.40	74.76	0.123	0.580	0.785
	$b = 100\%$	83.62	85.55	-0.006	0.518	0.822

Models constructed on data sets with an increasing proportion of elements pseudo-randomized in the response variable, showed increasing oob errors (Table 7.4): an increase of 1.45%, 9.93% and 55.59%, with 1%, 10% and 100% of the elements pseudo-randomized, respectively. The test set error values revealed that model performances did deteriorate gradually with increasing percentages of the elements pseudo-randomized. For the other evaluation statistics, similar conclusions hold. This result stresses the importance of accurate species mapping and vegetation type determination. The better model performances with 1% of the elements pseudo-randomized indicate that vegetation clustering uncertainty is prevalent in the reference data.

A possible way to get rid of the uncertainty associated with species clustering is to use a selection of dominant species instead of vegetation types for distribution modelling [51]. However, since vegetation types are frequently used in nature conservation, management and legislation (e.g. [295–298]), the application of vegetation distribution models will remain important.

7.5 Conclusions

Vegetation distribution models tend to describe vegetation patterns based on environmental variables. A variety of uncertainty sources can affect vegetation distribution modelling results. We investigated two of them; namely the uncertainties associated with (i) the spatial interpolation of environmental variables, and with (ii) species clustering into vegetation types. The following conclusions could be drawn from this investigation:

1. Variation partitioning is a useful methodology to assess importance and relevance of environmental variables and their spatial structuring to vegetation distributions. In this study, variation partitioning of the observed data (species, environmental and spatial) stressed the importance to include the important environmental variables constraining the vegetation distribution, as well as their spatial variability within the study area, for distribution modelling.

2. The sequential Gaussian simulation algorithm (which preserves data roughness and allows for local uncertainty assessment) is appropriate to simulate the spatial distributions of environmental variables based on point observations. Its ability to quantify local uncertainty is advantageous for the interpretation of distribution modelling results. In this study, simulation results were not accurate for most of the environmental variables, and conditional cumulative density functions showed a high variability for most grid cells.

3. The random forest distribution model generates an ensemble of classifiers, allowing for model output uncertainty assessment which could be quantified using Shannon's entropy measure.

4. The uncertainty associated with spatial interpolation of environmental variables propagated clearly through the distribution model and resulted in deteriorating model performances (higher error and H values, lower κ and $\text{AUC}_{\text{total}}$ values). Environmental variables with low uncertainty are primordial for adequate distribution modelling.

5. Pseudo-randomization tests are appropriate for uncertainty assessment associated with species clustering into vegetation types. Model performances on pseudo-randomized test sets emphasized the importance of accurate species clustering.

7.6 Additional remark on the use of time series

Ecological distribution modelling studies frequently rely on a limited number of model input variables to describe environmental space. Furthermore, the environmental gradients described by these variables are frequently changing throughout time (non-equilibrium situation, Chapter 1). Nevertheless, these dynamics are often ignored –mostly because of the costly and time-consuming monitoring work which is required to capture them– and environmental input data rely on a single measurement, or they are approximated by averages based on a limited number of observations. Obviously, uncertainty is introduced by such practice.

In this study, the environmental space was described by means of several variables related with groundwater quantity and dynamics, groundwater quality, soil and management. The temporal resolution of the observations differed for the different variables: groundwater dynamics were monitored every two weeks during the monitoring period, groundwater quality observations were only made twice, soil samples were only taken once and the different management regimes were delineated once (Table 7.1). Groundwater quantity, groundwater quality and soil measurements were used to derive the observational data sets $\{z(\mathbf{v}_\alpha), \alpha = 1, \ldots, n\}$ of the continuous environmental variables from which the spatial interpolation started.

Up to three different sources of uncertainty could be attributed to the different steps in the derivation of these observational data sets. The first source of uncertainty is the uncertainty associated with the individual measurements in the field (measurement error) of an environmental variable Z on time step i, $u(z_i)$. Based on these (uncertain) observations, a time series $\hat{\mathbf{z}} = (z_1, \ldots, z_t)$, with t the total number of time steps on which observations were made, was constructed. A second source of uncertainty is introduced by averaging the time series $\hat{\mathbf{z}}$. The uncertainty of the average value $u(\bar{\hat{\mathbf{z}}})$, assuming unbiased measurements and value independent measurement errors, can be expressed in a discrete way as [299]:

$$u(\bar{\hat{\mathbf{z}}}) = \sqrt{\frac{\sum_{i=1}^{t} u^2(z_i)}{t}}\,, \tag{7.6}$$

where $u(z_i)$ is the uncertainty on an individual measurement z_i, which is an element of the time series $\hat{\mathbf{z}}$ with a total of t measurements of variable Z. A third source of uncertainty is the uncertainty due to incomplete time coverage. This uncertainty is strongly affected by the frequency distribution of field measurements and the temporal occurrence of data gaps in the time series. A simplistic approach to assess the uncertainty on the average value due to incomplete time coverage, $u_{\mathrm{TC}}(\bar{\hat{\mathbf{z}}})$ assumes a random distribution of missing values throughout the time period considered, and is given by [299]:

$$u_{\mathrm{TC}}(\bar{\hat{\mathbf{z}}}) = \sigma(\mathbf{z})\sqrt{\frac{1 - t/T}{t}}\,, \tag{7.7}$$

where t is the number time steps on which observations in the incomplete time series ($\hat{\mathbf{z}}$) were made and T the number number time steps on which observations should be made to construct an complete time series ($\mathbf{z}$) covering the same period. $\sigma(\mathbf{z})$ is the standard deviation associated with the entire data set, and can be calculated as:

$$\sigma(\mathbf{z}) = \sqrt{\frac{1}{T-1}\sum_{j=1}^{T}(\mathbf{z}_j - \mu(\mathbf{z}))^2}\,. \tag{7.8}$$

In this study, these three types of uncertainty are clearly present. The determination of groundwater quantity variables, average groundwater depth and amplitude of the groundwater depth is based on groundwater depth measurements. Uncertainty in the individual measurements of groundwater depth can be caused, for example, by a coarse or incorrect determination of the ground surface, or by erroneous recordings due malfunctioning of the monitoring device. Furthermore, time series of groundwater depth measurements are averaged to obtain the average groundwater depth. A final source of uncertainty is related to an incomplete time coverage of the groundwater depth measurements which were made every fortnight.

Assume a constant uncertainty of the individual groundwater depth measurements ($u(z_i)$, with z_i a groundwater depth measurement) of 0.02 m, then the uncertainty related with the averaging of the time series of groundwater depth measurements ($u(\bar{\hat{\mathbf{z}}})$) equals 0.02 m (Eq. (7.6)). The uncertainty resulting from incomplete time coverage was calculated for each piezometer separately, since it is proportional to the standard deviation of the –not measured– entire time series $\sigma(\mathbf{z})$. Based on the incomplete time series measured at the study area $\sigma(\hat{\mathbf{z}})$ was used as an estimation of the standard deviation, which ranged between 0.1 m and 0.6 m. Time series recorded at piezometers installed in terrain with heavily fluctuating groundwater depths had higher standard deviations. These sites were situated on the levee of the river Dijle (western border of the study area), and fluctuations gradually decreased toward the central depression, to be followed by a slight increase at the eastern border. Fig. 7.11 visualizes the decline of $u_{\mathrm{TC}}(\bar{\hat{\mathbf{z}}})$ in function of percentage of time covered ($t/T \times 100$) for $\sigma(\hat{\mathbf{z}})$ values of 0.1 m, 0.3 m and 0.6 m.

The uncertainty associated with incomplete time coverage ranged from 0 when $t/T = 1$ (complete time coverage) to ∞ when $t/T = 0$ (fully incomplete time coverage). The effect of the standard deviation on $u_{\mathrm{TC}}(\bar{\hat{\mathbf{z}}})$ could clearly be seen. For piezometers where groundwater depth measurements did not change a lot throughout the year uncertainties are small compared to those where high fluctuations occurred. The interpretation is straightforward; time-incomplete measurements made at sites with limited groundwater dynamics are able to capture a bigger proportion of information compared with incomplete measurements made on highly

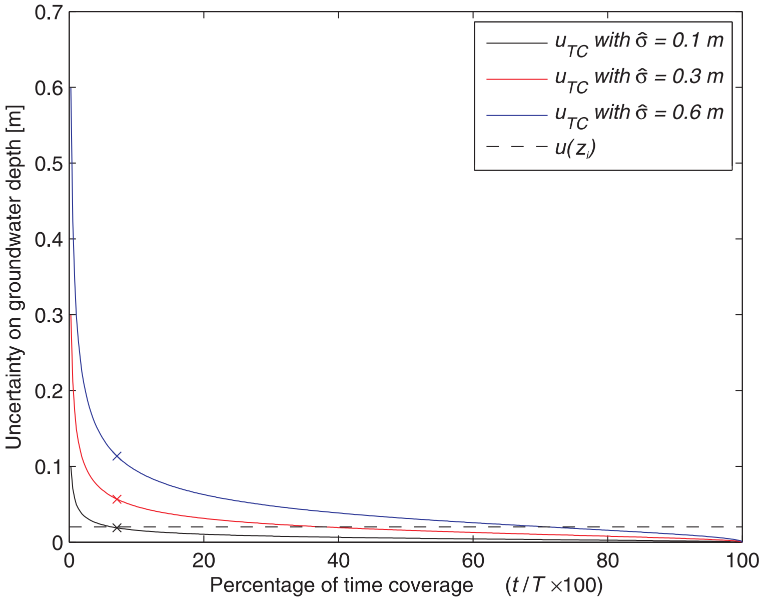


FIGURE 7.11 – The uncertainty of the individual groundwater depth measurements ($u(z_i)$) and uncertainty in function of incomplete time coverage for sites with different groundwater depth dynamics (represented by standard deviations $\sigma(\hat{\mathbf{z}})$).

dynamic sites. Therefore, the dynamics of the measured variable has a major influence on the uncertainty related with its incomplete measurement.

In this study where groundwater depth recordings were taken every two weeks during a 3 year period ($t = 26$ in 1991, 1992 and 1993), on an ideal measurement frequency of once per day (T = 365 in 1991 and 1993, and T = 366 in 1992), $u_{\mathrm{TC}}(\bar{\hat{\mathbf{z}}})$ values have to be read according $t/T \times 100 = 7$ (Fig. 7.11), and they range between 0.019 m for $\sigma(\hat{\mathbf{z}}) = 0.1$ and 0.11 m for $\sigma(\hat{\mathbf{z}}) = 0.6$. Uncertainties due to incomplete time coverage are almost equal to uncertainties related with the individual measurements at sites with low groundwater dynamics, while they are more than 5 times as large in highly dynamical sites.

For the amplitude of the groundwater depth, a similar uncertainty decomposition holds, with as only difference that there is no uncertainty due to averaging. The amplitude is calculated as the difference between minimal and maximal groundwater depths, and uncertainties associated with the individual measurements of these dates equal 0.02 m. The uncertainty on the groundwater amplitude caused by measurement uncertainty is therefore 0.028 m ($=\sqrt{(0.02)^2 + (0.02)^2}$). The determination of the amplitude of the groundwater depth is based on the same groundwater depth measurements as above, and uncertainty values at the different piezometers due to incomplete time coverage of the groundwater depth measurements are therefore the same.

The groundwater quality was assessed by concentration measurements of several hydrochemical groundwater compounds, which were only sampled twice, in spring and autumn of 1993, giving the wrong impression that they would be relatively static. This is not the case, as reported in numerous studies for various groundwater nutrients (e.g. [300–304]) and Fig. 1.3 where changes in nutrient concentrations are given after soil submergence. Unfortunately, a reasonable estimation of the uncertainties associated with groundwater quality variables was not feasible here; the uncertainty on the individual measurements would be highly influenced by the accuracy of the analyzing methods, but other elements like sample transport and treatment can be important as well, but are hard to capture quantitatively in terms of uncertainty. Additionally, the very incomplete time coverage did not allow to make acceptable estimations of the standard deviations of the fluctuation in groundwater nutrient concentrations, and therefore the uncertainty associated with incomplete time coverage could not be determined.

This remark on the use of time series in ecohydrological studies, adds to one of the main findings of this chapter. The uncertainty associated with spatial interpolation of environmental variables propagates through the distribution model, but additionally, the uncertainty associated with the use of (error containing and incomplete) time series to characterize dynamical ecosystem properties will also influence the model predictions.

8

Conclusions and recommendations

8.1 Starting point of this dissertation

Wetlands can be ranked amongst the most highly threatened ecosystems on the planet [305]. Unfortunately the degradation and loss of wetlands are continuing, mainly due to drainage for agriculture, settlements and urbanization, and pollution. Although no accurate record of worldwide wetland losses have been kept [1], an estimated global loss of 50% since 1900 is reported frequently [305, 306]. During the first half of the previous century, this mostly occurred in the northern temperate zones, however since the 1950s, tropical and sub-tropical wetlands have also been disappearing rapidly [305].

However, a growing awareness of wetland functions and a recognition of its values resulted gradually in wetland protection. At first instance by means of national laws and agreements, and more recently, after recognizing cross-boundary wetland values, by means of international cooperation in wetland protection. The most significant intergovernmental cooperation on wetland conservation is the Convention on Wetlands of International Importance especially as Waterfowl Habitat, held in Ramsar, Iran, in 1971 [307]. The Ramsar Convention provides the framework for national action and international cooperation for the conservation and wise use of wetlands and their resources. The treaty adopted by the contracting parties states

> *"the conservation and wise use of all wetlands through local, regional and national actions and international cooperation, as a contribution*

> *towards achieving sustainable development throughout the world"*

and comprehends four commitments [307]:

- to designate at least one wetland at the time of accession for inclusion in the List of Wetlands of International Importance and to promote its conservation, and in addition to continue to designate suitable wetlands within its territory for the List of Wetlands of International Importance (Article 2.1);
- to include wetland conservation considerations in their national land-use planning. Member parties have committed themselves to formulate and implement this planning so as to promote, as far as possible, the wise use of wetlands in their territory (Article 3.1);
- contracting parties shall establish nature reserves at wetlands (Article 4); and
- contracting parties have also agreed to consult with other contracting parties about implementation of the Convention, especially in regard to transboundary wetlands, shared water systems, and shared species (Article 5).

This international framework on wetland conservation and their wise use through the implementation of actions at local up to international scale, stimulates wetland research. One important research topic entails the modelling of vegetation patterns based on abiotical environmental conditions in distribution modelling. This is highly relevant for wetland management and conservation by ultimately enabling the prediction of vegetation response on environmental changes and the definition of environmental conditions to obtain certain goal ecosystems. The application of ecohydrological distribution models has been the starting point of this dissertation, on which the research objectives were defined.

8.2 Answers to the research questions

The research objectives of this dissertation were (Chapter 2):

- The introduction of ensemble learning by applying the so called '*random forest*' technique in vegetation distribution modelling by development of a *random forest distribution model* for the prediction of wetland vegetation distributions based on environmental wetland conditions;
- The assessment of the *predictive ability* of the random forest distribution model;
- The *identification of important environmental variables* determining the wetland vegetation distribution by the random forest distribution model;

- The assessment of the *generalization ability* of the random forest distribution model; and
- The analysis of input *uncertainty propagation* through the random forest distribution model.

Eight research questions have been asked to meet the research objectives. Answers to these questions were given throughout this dissertation and reformulated here in a comprehensive way.

1. Which techniques are most frequently applied in distribution modelling?
The conceptual considerations given in Chapter 4 summarized different distribution modelling approaches. The majority of distribution models, however, are based on field observations which are used for empirical distribution modelling assuming an equilibrium state. A literature overview of techniques used in these models indicated generalized linear models and tree-based techniques as the most frequently applied.

2. Can the random forest technique be used for vegetation distribution modelling?
This research question covers different aspects. Therefore it was split in several subquestions:

(a) *Are there requirements concerning data format?* A common data format in distribution modelling is a combination of continuous and categorical variables describing the environment as independent variables and a binomial or multinomial response. As learnt from Chapter 4, the random forest technique can readily be applied on these data.

(b) *Is the model output meaningful within a distribution modelling context?* The probability of occurrence is generated as model output. Probability values range between 0 and 1, and could be interpreted as habitat suitability values within a vegetation distribution modelling context.

(c) *Can the model output be introduced into geographical information systems?* Based on spatially explicit environmental gradients the model generates spatially explicit maps indicating the probability of occurrence for one or more vegetation types. These maps are interpreted as habitat suitability maps.

The positive answer to previous subquestions allows to conclude that the random forest technique is suitable to be applied in distribution modelling.

3. Is the predictive ability of the random forest model satisfactorily?
In Chapter 4 a random forest distribution model was constructed on the ecohydrological data set. Modelling results were compared with those of a logistic regression model constructed on the same data set. A variety of evaluation statistics were used and indicated significantly better model performances of the random forest

model. Furthermore, a random forest distribution model with reduced complexity outperformed the logistic regression model (Chapter 5). Therefore the answer to this research question is affirmative, and the random forest technique could be concluded to lead to better predictive ecohydrological distribution models.

4. Can the random forest distribution model provide information concerning the importance of environmental variables constraining the vegetation distribution? The random forest algorithm provides an estimator for variable importances. This measure was used in Chapter 5 to rank environmental variables constraining the wetland vegetation according to their importance on vegetation distributions. Two subquestions were asked to further assess the possibilities of this estimator:

(a) *Would other techniques identify the same environmental variables as being important?* In comparison with ordination results, similar rankings were formulated. Nevertheless, a bias in the importance value of categorical variables was prevalent and the variable importance estimates should thus be used with care. Variable rankings for the individual vegetation types only showed a significant similarity for three out of seven vegetation types when rankings formulated by hierarchical partitioning of logistic regression models and by the variable importance measure within random forests were compared.

(b) *Is it possible to construct accurate random forest distribution models on a reduced data set, only including the most important environmental variables?* Random forest distribution models with decreasing complexity were constructed in Chapter 5. Stable model performances were observed for models using more than five important predictive variables, and a sharp increase in error was observed when complexity further decreased. For the case study presented, a minimum of six predictive variables had to be included in order to gain an accurate fit. This model was more accurate than a logistic regression model using almost three times as many predictive variables!

5. Is there a spatial trend in the random forest distribution modelling results? Chapter 6 revealed decreasing probability values toward the boundary areas between adjacent vegetation types. An inverse proportionality between predicted probabilities and the species similarity of vegetation types was attributed to the classification performances of the individual classifiers within the random forest model. Grid cells within the boundary area between similar vegetation types (high similarity) were classified as similar, but different vegetation types by comparable numbers of classifiers (low maximal probability value). In addition, species clustering for vegetation type determination of boundary grid cells, is likely to contain a higher uncertainty, which may be reflected in the modelling results (Chapter 7).

Furthermore, isolated grid cells and small areas surrounded by another vegetation type were frequently incorrectly classified and considered to be a weakness of the random forest distribution model concerning prediction accuracy. Based on results of Chapter 7, it may be concluded that a smoothening of the environmental gradients, due to, for example, an insufficient monitoring density or an inappropriate interpolation technique not fully capturing the local variability, could cause the inferior model performances for isolated grid cells and small areas.

6. Does a random forest distribution model, constructed on a given wetland, perform satisfactorily when tested on a similar but distant wetland?
A random forest constructed on data of one site was applied to data of a distant, though ecologically similar site in Chapter 6. Model evaluation proved an unsatisfactorily model performance and stressed a confinement consequential to the empirical nature of the distribution model.

7. Does the use of an ensemble modelling technique allow for uncertainty assessment?
The random forest model output is a discrete probability distribution over all response vegetation types. The uniformity of the discrete probability distribution allows to gain information on model output uncertainty, for example by means of an entropy measure as implemented in Chapter 7. An uncertainty assessment of this kind could not be made for single classifier distribution models.

8. How does input uncertainty propagate through the random forest distribution model?
A propagation of input uncertainty to modelling results was obvious from the assessment performed in Chapter 7. The discrete probability distribution over all response vegetation classes flattened and the model did not find good evidence to classify a grid cell as a certain vegetation type when input uncertainty was prevalent. Therefore, the use of variables with limited uncertainty was concluded to be primordial for adequate distribution modelling.

8.3 Contribution of this dissertation

The main contribution of this dissertation to ecohydrological research is the introduction of a new technique for vegetation distribution modelling based on environmental conditions leading toward better model performances. Additionally, the assessment of model complexity reduction only to include the most important environmental gradients constraining wetland vegetation distributions, and the development of a framework for uncertainty assessment adds to the practical applicability and enhances the interpretation of the modelling results.

8.4 Future perspectives

Although a new modelling technique was introduced, the conceptual modelling approach remained unchanged and the random forest distribution model predicted vegetation type distributions empirically, based on field observations assuming an equilibrium state. As a consequence, the model described a real wetland situation with a high precision, but sacrificed generality (Fig. 4.1) which made a successful model implementation impossible beyond the local conditions where it was constructed upon (conclusion of Chapter 6). This limitation defines important future research perspectives.

8.4.1 Compilation of data sets

The compilation of an extensive data set prospects an enhanced generality. Recalling Fig. 6.5, in which a conceptual representation is made of the realized niches of *Calthion palustris* at Doode Bemde and Snoekengracht. Merging both areas would define a larger proportion of the fundamental niche. Adding more sites would increase this proportion further, until the (hypothetical) survey in which all sites with *Calthion palustris* are included. The realized niche delimited as such, would include all environmental conditions in which the vegetation type under consideration flourishes. However, it should be kept in mind that intraspecific and interspecific interactions are artificially (not explicitly) included in this survey. A distribution model constructed on this data set would have higher generality. Nevertheless, vegetation occurrence predictions following, for example, environmental changes will contain errors due to an inexplicit incorporation of biotic processes in the model.

Despite this drawback, empirical distribution modelling seems to be the most successful option today. The scientific understanding of wetland processes in relation to vegetation type occurrences is too limited for a more mechanistic modelling approach. Hence, the compilation of an extensive data set is of general interest for high quality modelling, which should be addressed in future research.

Based on this dissertation, recommendations are made concerning data gathering in wetlands. A first recommendation emerges from the fact that accurate random forest models could be constructed on a limited number of environmental gradients. Consequently, data gathering through the monitoring of wetlands should only encompass a limited number of environmental gradients, preferably those with a more causal effect on vegetation occurrences (see Chapter 5), that are measured at a satisfying spatio-temporal resolution. As such, model input uncertainties can be reduced and model output would have a higher reliability. A second recommendation regards the modelled entity. In this dissertation, vegetation types were used, other studies used plant species. The advantage of vegetation distribution models is that the model outcome can be thought of as a (more or less) discrete spatial entity with a certain species composition. The disadvantages, on the other hand, are that it is not the vegetation as such, but the individuals composing the vegetation that are related to the environmental conditions. Environmental changes will affect the individual species in the first place, with subsequent vegetation alterations. Additionally, as demonstrated in Chapter 6, species clustering into vegetation types may induce additional uncertainty that is propagated to the modelling results. A potential strategy to overcome these disadvantages is to use a selection of dominant plant species (e.g. based on coverage, biomass, phytosociological aspects) upon which the modelling is based.

8.4.2 Time dependence

The majority of distribution models are stationary (e.g. [78, 97, 98, 100, 114, 125–131]), assuming a state of equilibrium, notwithstanding their ultimate goal to predict vegetation responses on environmental changes. A simple example is presented for clarification (Fig. 8.1). Consider a unimodal response of probability of occurrence of two given vegetation types (or species) in relation to an environmental gradient. A current (t_0) environmental state (A) at a particular location within a wetland results in a probability of occurrence value, $P_{A,1}$ and $P_{A,2}$, for both vegetation types, respectively. However, a given anthropogenic disturbance alters the current environmental state (A) to another state (B) during a time interval ($t_1 - t_0$), and a random forest distribution model is used to determine the probability of occurrence of the given vegetation types under these new conditions. The stationary model will predict a probability of occurrence ($P_{B,1}$ and $P_{B,2}$) for both vegetation types, under the assumption of ecosystem equilibrium. The vegetation type with the highest probability of occurrence will be the predicted one. The dotted line in Fig. 8.1 represents the equal probability line ($P_{C,1} = P_{C,2}$), which is modelled based on an environmental state C, and the application of the decision rule leads to the prediction of vegetation type 1 when the environmental gradient has a value lower than C, and vegetation type 2 when than value is higher than C. Consequently, the modelled vegetation response is entirely dependent on the environmental changes (no time lag between environmental change and vegetation response since an equilibrium is assumed at all time), and information on the duration of the establishment of the new vegetation type is lacking. As seen in Fig. 8.1, different trajectories to reach the new equilibrium are possible: (tr_1) a fast response of the vegetation where the new equilibrium is reached at t_1, (tr_2) a gradual change in vegetation lagging behind the environmental change and reaching equilibrium at t_3, and (tr_3) no response until t_2 whereupon a change in vegetation is observed to reach equilibrium at t_3.

The inclusion of time into distribution models is a challenging future research perspective, which would increase their value for conservation and restoration ap-

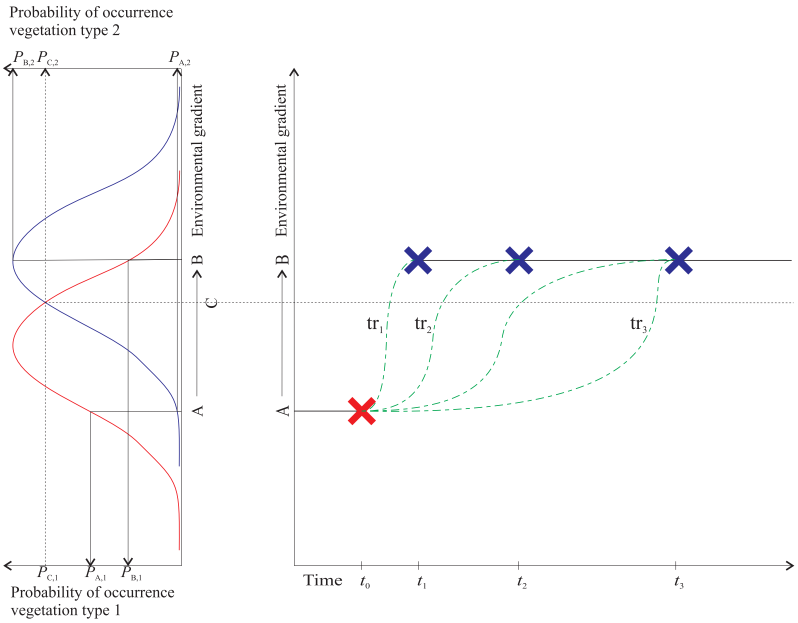


FIGURE 8.1 – Conceptual representation of the time dependence of vegetation responses to environmental changes. Two vegetation types have unimodal responses to an environmental gradient, showing an optimum, however, at different environmental states. An environmental state shift from A to B results in a decrease of the modelled probability of occurrence for vegetation type 1 ($P_{A,1} \rightarrow P_{B,1}$), while it increases for vegetation type 2 ($P_{A,2} \rightarrow P_{B,2}$). Predictions of a stationary model (crosses in right panel) do not give any information about the time-scale on which vegetation changes are occurring in response to change in environmental change. Different trajectories are possible (tr$_1$, tr$_2$ and tr$_3$, indicated by the green dashed line).

plications drastically. An empirical approach might justify the use of Markov models to simulate the transition dynamics of vegetation among different discrete vegetation types [308–310] (Clementsian approach). Stationary Markov models are based on a transition matrix containing probabilities of vegetation changes from one type to another over time. A stationary Markov model is entirely defined by the vegetation distribution at a given moment in time, and the transition probability between the different vegetation types. This seemingly simple model, however, becomes more complex when other processes such as immigration and extinction of (native and invasive) species [311] and spatial interactions [312] and gradients [313] have to be incorporated. Non-stationarity of the transition probabilities, which has been reported for natural ecosystems after major environmental perturbations [314], led to an extension of stationary Markov models to Markov-set models, using transition probability intervals [315], and hidden Markov models, accounting for additional processes overlaying the Markov process [316].

Nevertheless, the random forest distribution model does not include information on the transition between vegetation types. A possible methodology to overcome this shortcoming, is by applying coupled modelling (Subsection 8.4.3) in which the random forest distribution model is coupled to a transient model for the environmental setting, and defines another future research perspective.

8.4.3 Coupled modelling

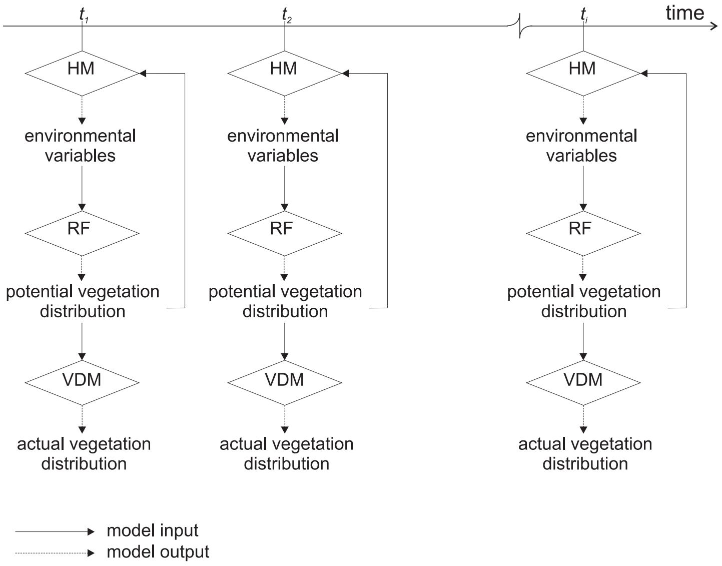
The idea behind a coupled modelling approach is to get time dependent and spatially distributed estimations for a range of environmental variables (with special attention for groundwater dynamics) by implementation of a hydrologic and/or hydrochemical model, on which the random forest vegetation distribution model is applied at several discrete time steps. Once again, the ability for vegetation distribution modelling based on a limited number of environmental variables (conclusion of Chapter 5) is advantageous since a decreased number of environmental variables are to be included, possibly decreasing the overall uncertainty on the environmental estimations at a given time, which is primordial for distribution modelling with a satisfactory accuracy level (conclusion of Chapter 7). A hydrologic and/or hydrochemical model for coupled modelling should satisfy following requirements:

1. the environmental estimations should be of appropriate time and spatial resolution;

2. the environmental estimations should be spatially distributed over the modelled area;

3. the model should account for all water fluxes to provide good estimates of quantitative and qualitative groundwater aspects.

A hydrologic modelling approach satisfying stated requirements could be the iterative modelling approach developed by Batelaan [79, 317], in which a spatially distributed water balance model (WetSpass, [318]) is connected to a regional groundwater model (MODFLOW, SEEPAGE [319] and DRAIN [320] packages) providing groundwater depth estimates. Input data for the WetSpass model are related to meteorology (precipitation, evapotranspiration, windspeed, temperature), land cover, slope, soil texture and groundwater depth to model the water balance of grids cells with high spatial resolution. Based on these environmental estimations, a random forest distribution model would be able to predict the probability of occurrence of several vegetation types at a given time, assuming a sufficiently wide environmental coverage in the data set where the distribution model was constructed upon (Fig. 8.2). The distribution model outcome should be interpreted as



t1
t2
ti
time
HM
environmental variables
RF
potential vegetation distribution
VDM
actual vegetation distribution
model input
model output

FIGURE 8.2 – Framework for coupled modelling.
Legend: HM = hydrologic or hydrochemical model, RF = random forest distribution model, and VDM = vegetation dynamics model.

an indicator for habitat suitability. For each vegetation type included, a spatially distributed suitability map shows the site potential given a set of modelled environmental constraints. Whether the vegetation type would appear following the environmental changes depends on localized immigration and extinction processes. Local immigration and extinction processes refer to individual-by-individual mortality and colonization, processes which are highly determined by the interaction with their neighbourhood. An additional model on vegetation or species dynamics actually accounting for neighbourhood interaction (e.g. spatial auto-correlation) could provide additional information on actual vegetation occurrences.

Additionally, hydrologic models may benefit as well from a coupled modelling approach. As the random forest distribution model predicts (potential) changes in vegetation distributions, associated physiological and structural changes of the vegetation and its dominant species might be taken into account. For example, in wetlands, where strong feedback from vegetation on site hydrology are immanent (see Fig 1.4), an important physiological alteration is related to the (local) immigration or extinction of phreatophytes at a given location. As demonstrated in numerous studies, phreatophytes consume water directly from the groundwater, resulting in a daily pattern of groundwater depth fluctuations [32, 321], while non-phreatophytes are restricted to the soil water in the vadose zone. Hence, the soil water balance (see Fig. 3.4) will differ when phreatophytes immigrate/extinct from a given wetland location. Similarly, the inclusion of structural vegetation changes may lead to better hydrologic modelling performances, by generating additional information for the calculation of water balances at a given location at a given time. Leaf area, for example, differs greatly among vegetation types [321]. Leaf area is proportional to the vegetation transpiration rates [322–324], and affects the energy and mass balance at the soil surface by shading [325, 326]. Therefore, feedback information from the random forest distribution model to the hydrologic model may effect improvement of the latter model.

A remark, however, should be made about the scale level on which the coupled modelling is performed, and its relative benefits. For modelling studies addressing hydrological balances at the local scale, coupled modelling would allow to gain more detailed insights by accounting for vegetation development and changes. For catchment modelling at the regional scale, however, a coupled modelling approach might be too elaborative and too data-demanding. Consider a forested catchment of a river with alluvial wetlands in its floodplain, as an example for clarification. A change from forest to agricultural land is foreseen, and the ecohydrological implications are under concern. For modelling the wetland area exclusively, a coupled modelling approach accounting for the hydrological and vegetation changes at the local scale would improve results. For modelling the entire catchment, however, vegetation changes within the wetland are of minor importance compared to the change from forest to agricultural land, and thus, a coupled modelling approach accounting for vegetation changes within the wetland would needlessly complicate the study.

In summary, a coupled modelling approach has potential for more detailed distribution modelling by taking response times on environmental changes (discretely) into account. However, future efforts regarding distribution model improvement should act commonly with additional high quality data acquisition since quality of data will eventually determine model performances and applicability.

8.5 Recommendations for potential users

Future perspectives (Section 8.4) were formulated from a theoretical point of view. However, the results of this dissertation allow to provide some practical recommendations for potential users of the random forest distribution model, with respect to monitoring and data acquisition, data preparation and the modelling itself.

Monitoring and data acquisition:

– Try to identify the environmental gradients that are constraining species or vegetation distributions mostly. Rather than monitoring a multitude of environmental variables on a low quality level, concentrate on a limited number of important ones (see Chapter 5).

- Some environmental gradients are easier to be described by categorical variables. Categorical variables are appropriate for random forest distribution modelling, a translation to dummy variables is unnecessary (see Chapter 4).
- The data quality of environmental variables and subsequent model performances are heavily influenced by the spatial and temporal monitoring resolution (see Chapter 7). Therefore, monitoring resolutions should be determined by the spatial and/or temporal dynamics of the variables under concern.
- If possible, focus on representative plant species rather than vegetation types. By doing so, the uncertainty introduced into the data by the clustering of species into vegetation types can be avoided (see Chapter 7). Representative plant species should be sufficiently specific to the environmental conditions. Rare species, however, may not be the best option since their presence/absence is determined to a relatively greater extent by other processes such as migration.
- In order to validate model performance, validation data should be acquired. Cross-validation is one possibility in which all data are used for both model construction and validation (see Chapter 4). Another possibility is the use of independent validation data. However, if the environmental gradients described by the validation data are not within the range of the data where the model was constructed upon, one can not expect satisfying modelling results (see Chapter 6). Therefore it is recommended to acquire independent validation data in the vicinity of the training data set.

Data preparation:

- Most likely a spatial interpolation of the acquired environmental variables has to be performed to get area covering estimates. Several geostatistical techniques are at one's disposal to perform this interpolation. The application of sequential Gaussian simulations (sGs), however, is recommended based on the results of chapter 5 because sGs preserves the characteristic roughness in the data, and has the ability to determine local uncertainty. Information on local uncertainty improves the interpretation of the random forest modelling results (e.g. are unsatisfactory modelling results at a certain locality related to uncertain environmental estimations?).
- The data were not normalized in this dissertation. It can be argued that data normalization could decrease the bias in the 'variable importance' measure (see Chapter 5) and improve model performances on independent data (see

chapter 6) because the environmental gradients of data would have a comparable range. Although independent modelling results are not expected to improve drastically, data normalization may be considered.

The random forest model:

- Based on chapter 7, a recommendation to look into the discrete probability distribution of the random forest model output, e.g. by means of Shannon's entropy, to get estimates on the uncertainty of the results is formulated. Doing so enhances the interpretation of the results a lot.

Bibliography

[1] W. J. Mitsch and J. G. Gosselink. *Wetlands*. Wiley & Sons, New York, 3rd edition, 2000.

[2] R. H. Kadlec and R. L. Knight. *Treatment Wetlands*. Lewis Publishers, Boca Raton, 1996.

[3] B. D. Wheeler. *Water and Plants in Freshwater Wetlands*. In A. J. Baird and R. L. Wilby, editors, Eco-hydrology: Plants and Water in Terrestrial and Aquatic Environments, pages 127–180. Routledge, London, 1999.

[4] G. Kirk. *The Biogeochemistry of Submerged Soils*. Wiley, Chichester, 2004.

[5] W. Armstrong. *Root Aeration in the Wetland Environment*. In D. D. Hook and R. M. M. Crawford, editors, Plant Life in Anaerobic Environments, pages 269–297. Ann Arbor Science, Woburn, MA, 1978.

[6] K. R. Reddy and E. M. D'Angelo. *Soil Processes Regulating Water Quality in Wetlands*. In W. Mitsch, editor, Global Wetlands: Old World and New, pages 309–324. Elsevier, Amsterdam, 1994.

[7] C. W. P. M. Blom. *Adaptations to flooding stress: from plant community to molecule*. Plant Biology, 1:261–273, 1999.

[8] M. B. Jackson and M. C. Drew. *Effects of Flooding on Growth and Metabolism of Herbaceous Plants*. In W. Mitsch, editor, Flooding and Plant Growth, pages 47–128. Academic Press, New York, 1984.

[9] M. C. Drew. *Oxygen deficiency and root metabolism: injury and acclimation under hypoxia and anoxia*. Annual Review of Plant Physiology and Plant Molecular Biology, 48:223–250, 1997.

[10] G. A. F. Hendry. *Oxygen and environmental stress in plants: an evolutionary context*. Proceedings of the Royal Society of Edinburgh, 102B:1–10, 1994.

[11] A. Rawyler, S. Arpgaus, and R. Breandle. *Impact of oxygen stress and energy availability on membrane stability of plant cells*. Annals of Botany, 90:499–507, 2002.

[12] G. T. Babcock. *How oxygen is activated and reduced in respiration*. Proceedings of the National Academy of Sciences of the USA, 96:12971–12973, 1999.

[13] F. Rébeillé, R. Bligny, and R. Douce. *Oxygen and temperature effects on the fatty acid composition in sycamore cells (Acer pseudoplatanus L.)*. Biochimica et Biophysica Acta, 620:1–9, 1980.

[14] E. J. W. Visser, L. A. C. J. Voesenek, B. B. Vartapetian, and M. B. Jackson. *Flooding and plant growth*. Annals of Botany, 91:107–109, 2003.

[15] C. W. P. M. Blom and L. A. C. J. Voesenek. *Flooding: the survival strategies of plants*. Trends in Ecology & Evolution, 11(7):290–295, 1996.

[16] P. Perata and A. Alpi. *Plant response to anaerobiosis*. Plant Science, 93:1–17, 1993.

[17] R. M. M. Crawford. *Studies in Plant Survival. Ecological Case Histories of Plant Adaptation to Adversity*. Blackwell, Studies in Ecology 11, 1989.

[18] O. A. Clevering, W. Van Vierssen, and C. W. P. M. Blom. *Growth, photosynthesis and carbohydrate utilization in submerged Scirpus maritimus L. during spring growth*. New Phytologist, 130:105–116, 1995.

[19] M. Kawase. *Anatomical and morphological adaptation of plants to waterlogging*. HortScience, 16(1):30–34, 1981.

[20] W. Armstrong, R. Brändle, and M. B. Jackson. *Mechanisms of flood tolerance in plants*. Acta Botanica Neerlandica, 43:307–358, 1994.

[21] W. M. H. G. Engelaar. *Root porosities and radial oxygen losses from Rumex and Plantago species as influenced by soil pore diameter and soil aeration*. New Phytologist, 125:565–574, 1993.

[22] T. T. Kozlowski. *Flooding and Plant Growth*. Academic Press, New York, 1984.

[23] C. W. P. M. Blom, G. M. Bögemann, P. Laan, A. J. M. Van der Sman, H. M. Van de Steeg, and L. A. C. J. Voesenek. *Adaptations to flooding in plants from river areas*. Aquatic Botany, 38:29–47, 1990.

[24] A. J. M. Van der Sman, L. A. C. J. Voesenek, C. W. P. M. Blom, F. J. M. Harren, and J. Reuss. *The role of ethylene in shoot elongation with respect to survival and seed output of flooded Rumex maritimus L. plants*. Functional Ecology, 5:304–313, 1991.

[25] N. Malmer, B. M. Svensson, and B. Wallen. *Interactions between Sphagnum mosses and field layer vascular plants in the development of peat forming systems*. Folia Geobotanica & Phytotaxonomica, 29(4):483–496, 1994.

[26] N. Vanbreemen. *How Sphagnum bogs down other plants*. Trends in Ecology & Evolution, 10(7):270–275, 1995.

[27] J. V. Ward and J. A. Stanford. *Ecological connectivity in alluvial river ecosystems and its disruption by flow regulation*. Regulated Rivers-Research and Management, 11(1):105–119, 1995.

[28] V. M. Castillo, M. Martinez Mena, and J. Albaladejo. *Runoff and soil loss response to vegetation removal in a semiarid environment*. Soil Science of America Journal, 61(4):1116–1121, 1994.

[29] R. Mainiero and M. Kazda. *Effects of Carex rostrata on soil oxygen in relation to soil moisture*. Plant and Soil, 270(1–2):311–320, 2005.

[30] I. Rodríguez-Iturbe and A. Porporato. *Ecohydrology of Water-Controlled Ecosystems*. Cambridge University Press, Cambridge, 2004.

[31] P. D'Odorico, K. Caylor, G. S. Okin, and T. M. Scanlon. *On soil moisture-vegetation feedbacks and their possible effects on the dynamics of dryland ecosystems*. Journal of Geophysical Research-Biogeosciences, 112(G4):G04010, 2007.

[32] V. Engel, E. G. Jobby, M. Steiglitz, M. Williams, and R. B. Jackson. *Hydrological consequences of Eucalyptus afforestation in the Argentine Pampas*. Water Resources Research, 41(10):W10409, 2005.

[33] L. Ridolfi L, P. D'Odorico, and F. Laio. *Effect of vegetation-water table feedbacks on the stability and resilience of plant ecosystems*. Water Resources Research, 42(1):W01201, 2006.

[34] K. A. M. Engelhardt and M. E. Ritchie. *Effects of macrophyte species richness on wetland ecosystem functioning and services*. Nature, 411:687–689, 2001.

[35] R. Costanza, R. d'Arge, R. de Groot, S. Farber, M. Grasso, B. Hannon, K. Limburg, S. Naeem, R. V. O'Neill, J. Paruelo, R. G. Raskin, P. Sutton, and M. van den Belt. *The value of the world's ecosystem services and natural capital*. Nature, 387:253–260, 1997.

[36] W. J. Mitsch and J. G. Gosselink. *The values of wetlands: importance of scale and landscape setting*. Ecological Economics, 35:25–33, 2000.

[37] J. C. Stromberg, R. Tiller, and B. Richter. *Effects of groundwater decline on riparian vegetation of semiarid regions: The San Pedro, Arizona*. Ecological Applications, 6(1):113–131, 1996.

[38] J. G. Gosselink and E. Maltby. *Wetland Losses and Gains*. In M. Williams, editor, Wetlands: A Threatened Landscape, pages 292–322. Basil Blackwell, Oxford, 1990.

[39] M. A. Saleque, U. A. Naher, A. Islam, A. B. M. B. U. Pathan, A. T. M. S. Hossain, and C. A. Meisner. *Inorganic and organic phosphorus fertilizer effects on the phosphorus fractionation in wetland rice soils*. Soil Science Society of America Journal, 68(5):1635–1644, 2004.

[40] E. P. H. Best. *The impact of mechanical harvesting regimes on the species composition of Dutch ditch vegetation - quantitative approach*. Journal of Aquatic Plant Management, 31:148–154, 1993.

[41] M. R. Gale, J. W. McLaughlin, M. F. Jurgensen, C. C. Trettin, T. Soelsepp, and P. O. Lydon. *Plant community responses to harvesting and post-harvest manipulations in a Picea-Larix-Pinus wetland with a mineral substrate*. Wetlands, 18(1):150–159, 1998.

[42] B. Blossey, L. C. Skinner, and J. Taylor. *Impact and management of purple loosestrife (Lythrum salicaria) in North America*. Biodiversity and Conservation, 10(10):1787–1807, 2001.

[43] J. E. Austin, J. R. Keough, and W. H. Pyle. *Effects of habitat management treatments on plant community composition and biomass in a montane wetland*. Wetlands, 27(3):570–587, 2007.

[44] J. T. Morris. *Effects of nitrogen loading on wetland ecosystems with particular reference to atmospheric deposition*. Annual Review of Ecology and Systematics, 22:257–279, 1991.

[45] B. L. Bedford, M. R. Walbridge, and A. Aldous. *Patterns in nutrient availability and plant diversity of temperate North American wetlands*. Ecology, 80(7):2151–2169, 1999.

[46] S. A. Levin. *The problem of pattern and scale in ecology*. Ecology, 73(6):1943–1967, 1992.

[47] E. J. Gustafson. *Quantifying landscape spatial pattern: what is the state of the art?* Ecosystems, 1:143–156, 1998.

[48] G. L. W. Perry. *Landscapes, space and equilibrium: shifting viewpoints*. Progress in Physical Geography, 26(3):339–359, 2002.

[49] B. schröder. *Pattern, process, and function in landscape ecology and catchment hydrology - how can quantitative landscape ecology support predictions in ungauged basins?* Hydrology and Earth System Sciences, 10:967–979, 2006.

[50] G. E. Hutchinson. *Concluding remarks*. Cold Spring Harbor Symposia on Quantitative Biology, 22(2):415–427, 1957.

[51] A. Guisan and N. E. Zimmerman. *Predictive habitat distribution models in ecology*. Ecological Modelling, 135(2–3):147–186, 2000.

[52] J. Wu and S. A. Levin. *A patch-based modeling approach: conceptual framework and simulation scheme*. Ecological Modelling, 101(2–3):325–346, 1997.

[53] D. Hannah, P. Wood, and J. Sadler. *Ecohydrology and hydroecology: A 'new paradigm'?* Hydrological Processes, 18:3439–3445, 2004.

[54] M. J. Wassen and A. P. Grootjans. *Ecohydrology: an interdisciplinary approach for wetlands management and restoration*. Vegetatio, 126:1–4, 1996.

[55] A. J. Baird and Robert L. Wilby, editors. *Eco-hydrology: Plants and Water in Terrestrial and Aquatic Environments*. Routledge, London, 1999.

[56] I. Rodriguez–Iturbe. *Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics*. Water Resources Research, 36:3–9, 2000.

[57] E. W. Seabloom, K. A. Moloney, and A. G. Van Der Valk. *Constraints on the establishment of plants along a fluctuating water-depth gradient*. Ecology, 82:2216–2232, 2001.

[58] V. A. Cramer and R. J. Hobbs. *Ecological consequences of altered hydrological regimes in fragmented ecosystems in southern Australia: impact and possible management responses*. Austral Ecology, 27:546–564, 2002.

[59] M. Zalewski, V. Santiago-Fandino, and J. Neate. *Energy, water, plant interactions: 'green feedback' as a mechanism for environmental management and control trough the application of phytotechnology and ecohydrology*. Hydrological Processes, 17:2753–2767, 2003.

[60] R. H. Kemmers, J. M. J. Gieske, P. Veen, and L. M. L. Zonneveld. *Standaard meetprotocol verdroging: voorlopige richtlijnen voor monitoring van anti-verdrogingsprojecten*. Rapport 15–1, Nationaal Onderzoeksprogramma Verdroging, 1995. (In Dutch).

[61] I. Somodi and Z. Botta-Dukat. *Determinants of floating island vegetation and succession in a recently flooded shallow lake, Kis-Balaton (Hungary)*. Aquatic Botany, 79:357–366, 2004.

[62] A. van Dijk. *Ecohydrology: it's all in the game?* Hydrological Processes, 18:3683–3686, 2004.

[63] J. L. Monteith. *Evaporation and the environment*. In The State and Movement of Water in Living Organisms, number 19, pages 205–234, Swansea, 1965. Symposia of the Society for Experimental Biology, University Press.

[64] I. Rodriguez-Iturbe, A. Porporato, F. Laio, and L. Ridolfi. *Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress I. Scope and general outline*. Advances in Water Resources, 24:695–705, 2001.

[65] F. Laio, A. Porporato, L. Ridolfi, and I. Rodriguez-Iturbe. *Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress II. Probabilistic soil moisture dynamics*. Advances in Water Resources, 24:707–723, 2001.

[66] F. Laio, A. Porporato, C. P. Fernandez-Illescas, and I. Rodriguez-Iturbe. *Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress IV. Discussion and real cases*. Advances in Water Resources, 24:745–762, 2001.

[67] A. Porporato, F. Laio, L. Ridolfi, and I. Rodriguez-Iturbe. *Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress III. Vegetation water stress*. Advances in Water Resources, 24:725–744, 2001.

[68] I. Rodriguez-Iturbe, P. D'Odorico, F. Laio, L. Ridolfi, and S. Tamea. *Challenges in humid land ecohydrology: Interactions of water table and unsaturated zone with climate, soil, and vegetation*. Water Resources Research, 43:W09301, 2006.

[69] W. Huybrechts, E. De Bie, P. De Becker, M. Wassen, and A. Bio. *Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, ITORS-VL (VLINA 00/16)*. Instituut voor Natuurbehoud, Brussel, 2002. (In Dutch).

[70] W. Huybrechts, O. Batelaan, P. De Becker, I. Joris, and P. van Rossum. *Ecohydrologisch Onderzoek Waterrijke Vallei-ecosystemen (VLINA 96/03)*. Instituut voor Natuurbehoud, Brussel, 2000. (In Dutch).

[71] P. De Becker and W. Huybrechts. *De Doode Bemde - Ecohydrologische Atlas*. Instituut voor Natuurbehoud, Brussel, 2000b. (In Dutch).

[72] W. Huybrechts and P. De Becker. *De Snoekengracht - Ecohydrologische Atlas*. Instituut voor Natuurbehoud, Brussel, 1999. (In Dutch).

[73] W. Huybrechts and P. De Becker. *Vorsdonkbos - Turfputten - Ecohydrologische Atlas*. Instituut voor Natuurbehoud, Brussel, 2000. (In Dutch).

[74] P. De Becker and W. Huybrechts. *Vallei van de Zwarte Beek - Ecohydrologische Atlas*. Instituut voor Natuurbehoud, Brussel, 2000a. (In Dutch).

[75] N. E. C. Verhoest, P. A. Troch, and F. A. De Troch. *On the applicability of Barlett-Lewis rectangular pulses models in the modeling of design storms at a point*. Journal of Hydrology, 202:108–120, 1997.

[76] I. L. M. De Jongh, N. E. C. Verhoest, and F. De Troch. *Analysis of a 105-year time series of precipitation observed at Uccle, Belgium*. International Journal of Climatology, 26:2023–2039, 2006.

[77] Y. Van Herpe and P. A. Troch. *Spatial and temporal variations in surface water nitrate concentrations in a mixed land use catchment under humid temperate climatic conditions*. Hydrological Processes, 14:2439–2455, 2000.

[78] A. M. F. Bio, P. De Becker, E. De Bie, W. Huybrechts, and M. Wassen. *Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions*. Biodiversity and Conservation, 11:2189–2216, 2002.

[79] O. Batelaan. *Phreatology. Characterizing groundwater recharge and discharge using remote sensing, GIS, ecology, hydrochemistry and groundwater modelling*. PhD thesis, Vrije Universiteit Brussel, April 2006.

[80] I. Joris and J. Feyen. *Modelling water flow and seasonal soil moisture dynamics in an alluvial groundwater-fed wetland*. Hydrology and Earth System Sciences, 7(1):57–66, 2003.

[81] P. De Becker, M. Hermy, and J. Butaye. *Ecohydrological characterisation of a groundwater-fed alluvial floodplane mire*. Applied Vegetation Science, 2:215–228, 1999.

[82] G. Londo. *Nederlandse Freatophyten*. Pudoc, Wageningen, 1988. (In Dutch).

[83] M. O. Hill. *TWINSPAN - a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes*. Cornell University, Ithaca, New York, 1979.

[84] I. Joris. *Soil water dynamics in alluvial wetlands: Field and modelling study*. PhD thesis, Katholieke Universiteit Leuven, April 2005.

[85] P. F. Hudson. *Natural Levees*. Encyclopedia of Water Science, doi: 10.1081/E-EWS-120038052:1–4, 2005.

[86] M. F. P. Bierkens. *Complex confining layers: A stochastic analysis of hydraulic properties at various scales*. PhD thesis, Universiteit Utrecht, 1994.

[87] D. R. Maidment, editor. *Handbook of Hydrology*. McGraw-Hill, New York, 1992.

[88] R. E. White. *Principles and Practice of Soil Science: The Soil as Natural Resource*. Blackwell Science, Oxford, 1997.

[89] A. K. Knapp, JJ. T. Fahnestock, S. P. Hamburg, L. B. Statland, T. R. Seastedt, and D. S. Schimel. *Landscape patterns in soil plant water relations and primary production in tallgrass prairie*. Ecology, 74(2):549–560, 1993.

[90] H. O. Venterink and M. J. Wassen. *A comparison of six models predicting vegetation response to hydrological habitat change*. Ecological Modelling, 101(2-3):347–361, 1997.

[91] A. C. D. Ertsen, A. M. F. Bio, W. Bleuten, and M. J. Wassen. *Comparison of the performance of species response models in several landscape units in the province of Noord-Holland, The Netherlands*. Ecological Modelling, 109(2):213–223, 1998.

[92] A. M. F. Bio, R. Alkemade, and A. Barendregt. *Determining alternative models for vegetation response analysis: a non-parametric approach*. Journal of Vegetation Science, 9:5–16, 1998.

[93] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, Boca Raton, 2nd edition, 1989.

[94] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[95] T. W. Yee and N. D. Mitchell. *Generalized additive models in plant ecology*. Journal of Vegetation Science, 2:587–602, 1991.

[96] N. H. Augustin, R. P. Cummins, and D. D. French. *Exploring spatial vegetation dynamics using logistic regression and a multinomial logit model*. Journal of Applied Ecology, 38:991–1006, 2001.

[97] M. P. Austin. *Spatial prediction of species distribution: an interface between ecological theory and statistical modeling*. Ecological Modelling, 157(2–3):101–118, 2002.

[98] R. Engler, A. Guisan, and L. Rechsteiner. *An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data*. Journal of Applied Ecology, 41:263–274, 2004.

[99] S. P. Rushton, S. J. Ormerod, and G. Kerby. *New paradigms for modelling species distributions*. Journal of Applied Ecology, 41:193–200, 2004.

[100] P. Segurado and M. B. Araújo. *An evaluation of methods for modelling species distributions*. Journal of Biogeography, 31:1555–1568, 2004.

[101] R. W. Fitzgerald. *The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain*. In American Society of Photogrammetry and Remote Sensing, Proceedings of the XVII Congress ASPRS, pages 570–573, Bethesda MD, 1992.

[102] F. Recknagel. *Applications of machine learning to ecological modelling*. Ecological Modelling, 146:303–310, 2001.

[103] M. O. Hill. *Reciprocal averaging - eigenvector method of ordination*. Journal of Ecology, 61:237–244, 1973.

[104] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

[105] H. S. Fischer. *Simulating the distribution of plant communities in an alpine landscape*. Coenoses, 5:37–49, 1990.

[106] E. O. Box. *Macroclimate and Plant Forms: An Introduction to Predictive Modeling in Phytogeography*. Junk, The Hague, 1992.

[107] R. Levins. *The stategy of model building in population biology*. American Scientist, 54:421–431, 1966.

[108] F. E. Clements. *Plant Succession: An Analysis of Development of Vegetation*. Carnegie Institution of Washington, Washington, 1916.

[109] H. A. Gleason. *The individualistic concept of plant association*. Bulletin of the Torrey Botany Club, 53:7–26, 1926.

[110] J. Nelder and R. Wedderburn. *Generalized linear models*. Journal of the Royal Statistical Society. Series A, 135:370–384, 1972.

[111] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, London, 2nd edition, 2001.

[112] J. Oksanen and P. R. Minchin. *Continuum theory revisited: what shape are species responses along ecological gradients?* Ecological Modelling, 157(2–3):119–129, 2002.

[113] W. Thuiller, M. B. Araújo, and S. Lavorel. *Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales*. Journal of Vegetation Science, 14(5):669–680, 2003.

[114] J. J. Lawler, D. White, R. P. Neilson, and A. R. Blaustein. *Predicting climate-induced range shifts: model differences and model reliability*. Global Change Biology, 12:1568–1584, 2006.

[115] R. Kay and S. Little. *Transformations of the explanatory variables in the logistic regression model for binary data*. Biometrika, 74(3):495–501, 1987.

[116] M. P. Austin, A. O. Nicholls, M. D. Doherty, and J. A. Meyers. *Determining species response functions to an environmental gradient by means of a β-function*. Journal of Vegetation Science, 5(2):215–228, 1994.

[117] J. Huisman, H. Olff, and L. F. M. Fresco. *A hierarchical set of models for species response analysis*. Journal of Vegetation Science, 4(1):37–46, 1993.

[118] J. L. Uzó-Domènech, J. Mateu, and J. A. Lopez. *Mathematical and statistical formulation of an ecological model with applications*. Ecological Modelling, 101(1):27–40, 1997.

[119] M. D. Jennings, J. W. Williams, and M. R. Stromberg. *Diversity and productivity of plant communities across the Inland Northwest, USA*. Oecologia, 143(4):607–618, 2005.

[120] J. L. Espinar. *Sample size and the detection of a hump-shaped relationship between biomass and species richness in Mediterranean wetlands*. Journal of Vegetation Science, 17(2):227–232, 2006.

[121] C. N. Meynard and J. F. Quinn. *Predicting species distributions: a critical comparison of the most common statistical models using artificial species*. Journal of Biogeography, 34:1455–1469, 2007.

[122] K. Wilson, B. T. Grenfell, and D. J. Shaw. *Analysis of aggregated parasite distributions: A comparison of methods*. Functional Ecology, 10(5):592–601, 1996.

[123] O. R. Vetaas. *The effect of canopy disturbance on species richness in a central Himalayan oak forest*. Plant Ecology, 132(1):29–38, 1997.

[124] A. G. Teira and B. Peco. *Modelling oldfield species richness in a mountain area*. Plant Ecology, 166(2):249–261, 2003.

[125] W. Thuiller. *BIOMOD — optimizing predictions of species distributions and projecting potential future shifts under global change*. Global Change Biology, 9(10):1353–1362, 2003.

[126] A. Guisan, J. P. Theurillat, and F. Kienast. *Predicting the potential distribution of plant species in an Alpine environment*. Journal of Vegetation Science, 9(1):65–74, 1998.

[127] A. Guisan, S. B. Weiss, and A. D. Weiss. *GLM versus CCA spatial modeling of plant species distribution*. Plant Ecology, 143(1):107–122, 1999.

[128] P. W. van Horssen, P. P. Schot, and A. Barendregt. *A GIS-based plant prediction model for wetland ecosystems*. Landscape Ecology, 14:253–265, 1999.

[129] J. Pearce and S. Ferrier. *An evaluation of alternative algorithms for fitting species distribution models using logistic regression*. Ecological Modelling, 128(3):127–147, 2000.

[130] J. Miller and J. Franklin. *Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence*. Ecological Modelling, 157(2–3):227–247, 2002.

[131] J. Peters, B. De Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts. *Random forests as a tool for predictive ecohydrological modelling*. Ecological Modelling, 207(2–4):304–318, 2007.

[132] M. Horsak. *Mollusc community patterns and species response curves along a mineral richness gradient: a case study in fens*. Journal of Biogeography, 33(1):98–107, 2006.

[133] M. P. Austin and J. A. Meyers. *Current approaches to modeling the environmental niche of eucalypts: implication for management of forest biodiversity*. Forest Ecology Management, 85:95–106, 1996.

[134] J.R. Leathwick and G.M. Rogers. *Modeling relationships between environmental and canopy composition in secondary vegetation in central North Island, New Zealand*. New Zealand Journal of Ecology, 20:147–161, 1996.

[135] J. Franklin. *Predicting the distribution of shrub species in southern California from climate and terrain-derived variables*. Journal of Vegetation Science, 9(5):733–748, 1998.

[136] A. Lehmann. *GIS modeling of submerged macrophyte distribution using generalized additive models*. Plant Ecology, 139:113–124, 1998.

[137] J. Pearce and S. Ferrier. *The practical value of modelling relative abundance of species for regional conservation planning: a case study*. Biological Conservation, 98:33–43, 2001.

[138] J. Pykälä, M. Luoto, R. K. Heikkinen, and T. Kontula. *Plant species richness and persistence of rare plants in abandoned semi-natural grasslands in northern Europe*. Basic and Applied Ecology, 6(1):25–33, 2005.

[139] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[140] J. M. Lenihan and R. P. Neilson. *A rule-based vegetation formation model for Canada*. Journal of Biogeography, 20:615–628, 1993.

[141] S. E. Franklin and B. A. Wilson. *Vegetation mapping and change detection using SPOT MLA and Landsat imagery in Kluane National Park*. Canadian Journal of Remote Sensing, 17:2–17, 1991.

[142] L. Torgo. *Functional models for regression tree leaves*. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 385–393, 1997.

[143] G. De'Ath. *Multivariate regression trees: a new technique for modeling species-environment relationships*. Ecology, 83(4):1105–1117, 2002.

[144] L. R. Iverson and A. M. Prasad. *Predicting abundance of 80 tree species following climate change in the eastern United States*. Ecological Monographs, 68(4):465–485, 1998.

[145] G. De'ath and K. E. Fabricius. *Classification and regression trees: A powerful yet simple technique for ecological data analysis*. Ecology, 81(11):3178–3192, 2000.

[146] G. G. Moisen and T. S. Frescino. *Comparing five modelling techniques for predicting forest characteristics*. Ecological Modelling, 157(2–3):209–225, 2002.

[147] Z. Lososova, M. Chytry, and I. Kuehn. *Plant attributes determining the regional abundance of weeds on central European arable land*. Journal of Biogeography, 35(1):177–187, 2008.

[148] R. L. Lawrence and W. J. Ripple. *Fifteen years of revegetation of Mount St. Helens: A landscape-scale analysis*. Ecology, 81(10):2742–2752, 2000.

[149] A. W. Crall, G. J. Newman, T. J. Stohlgren, C. S. Jarnevich, P. Evangelista, and D. Geuenther. *Evaluating dominance as a component of non-native species invasions*. Diversity and Distributions, 12(2):195–204, 2006.

[150] M. Rouget, D. M. Richardson, R. M. Cowling RM, J. W. Lloyd, and A. T. Lombard. *Current patterns of habitat transformation and future threats to biodiversity in terrestrial ecosystems of the Cape Floristic Region, South Africa*. Biological Conservation, 112(1–2):63–85, 2003.

[151] T. C. Edwards Jr., D. R. Cutler, N. E. Zimmermann, L. Geiser, and G. G. Moisen. *Effects of sample survey design on the accuracy of classification tree models in species distribution models*. Ecological Modelling, 199(2):132–141, 2006.

[152] P. Segurado, M. B. Araújo, and W. E. Kunin. *Consequences of spatial autocorrelation for niche-based models*. Journal of Applied Ecology, 43(3):433–444, 2006.

[153] D. M. Cairns. *A comparison of methods for predicting vegetation type*. Plant Ecology, 156(1):3–18, 2001.

[154] J. Franklin. *Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral*. Applied Vegetation Science, 5(1):135–146, 2002.

[155] T. C. Edwards, D. R. Cutler, N. E. Zimmerman, L. Geiser, and J. Alegria. *Model-based stratifications for enhancing the detection of rare ecoligical events*. Ecology, 86(5):1081–1090, 2005.

[156] J. Miller and J. Franklin. *Explicitly incorporating spatial dependence in predictive vegetation models in the form of explanatory variables: a Mojave Desert case study*. Journal of Geographical Systems, 8(4):411–435, 2006.

[157] M. Lindbladh, G. L. Jacobson, and M. Schauffler. *The postglacial history of three Picea species in New England, USA*. Quaternary Research, 59(1):61–69, 2003.

[158] T. Matsui, T. Yagihashi, T. Nakaya, N. Tanaka, and H. Taoda. *Climatic controls on distribution of Fagus crenata forests in Japan*. Journal of Vegetation Science, 15(1):57–66, 2004.

[159] A. Accad and D. T. Neil. *Modelling pre-clearing vegetation distribution using GIS-integrated statistical, ecological and data models: A case study*

from the wet tropics of Northeastern Australia. Ecological Modelling, 198(1–2):85–100, 2006.

[160] W. Thuiller, J. Vayreda, J. Pino, S. Sabate, S. Lavorel, and C. Gracia. *Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain)*. Global Ecology and Biogeography, 12(4):313–325, 2003.

[161] M. B. Araújo, R. G. Pearson, W. Thuiller, and M. Erhard. *Validation of species-climate impact models under climate change*. Global Change Biology, 11(9):1504–1513, 2005.

[162] L. Gustafsson and I. Eriksson. *Factors of importance for epiphytic vegetation of aspen Populus-tremula with special emphasis on bark chemistry and soil chemistry*. Journal of Applied Ecology, 32(2):412–424, 1995.

[163] J. L. Ohmann and T. A. Spies. *Regional gradient analysis and spatial pattern of woody plant communities of Oregon forests*. Ecological Monographs, 68(2):151–182, 1998.

[164] R. H. Okland, K. Rydgren, and T. Okland. *Single-tree influence on understorey vegetation in a Norwegian boreal spruce forest*. Oikos, 87(3):488–498, 1999.

[165] A. T. Oliveirafilho, E. A. Vilela, M. L. Gavilanes, and D. A. Carvalho. *Effect of flooding regime and understorey bamboos on the physiognomy and tree species composition of a tropical semideciduous forest in southeast Brazil*. Vegetatio, 113(2):99–124, 1994.

[166] A. Rubio and A. Escudero. *Small-scale spatial soil-plant relationship in semi-arid gypsum environments*. Plant and Soil, 220(1–2):139–150, 2000.

[167] D. J. Eldridge and M. E. Tozer. *Environmental factors relating to the distribution of terricolous bryophytes and lichens in semi-arid eastern Australia*. Bryologist, 100(1):28–39, 1997.

[168] S. Deblois and A. Bouchard. *Dynamics of Thuja-Occidentalis in an agricultural landscape of southern Quebec*. Journal of Vegetation Science, 6(4):531–542, 1995.

[169] M. O. Hill. *Patterns of species distribution in Britain elucidated by canonical correspondence analysis*. Journal of Biogeography, 18:247–255, 1991.

[170] M. Gottfried, H. Pauli, and G. Grabherr. *Predictions of vegetation patterns at the limits of plant life: a new view of the alpine-nival ecotones*. Arctic and Alpine Research, 30:207–221, 1998.

[171] P. Jeanneret, B. Schupbach, and H. Luka. *Quantifying the impact of landscape and habitat features on biodiversity in cultivated landscapes*. Agriculture Ecosystems & Environment, 98(1–3):311–320, 2003.

[172] P. Legendre and M. J. Anderson. *Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments*. Ecological Monographs, 69(1):1–24, 1999.

[173] P. Legendre and E. D. Gallegher. *Ecologically meaningful transformations for ordination of species data*. Oecologia, 129(2):271–280, 2001.

[174] J. Leps. *Nutrient status, disturbance and competition: an experimental test of relationships in a wet meadow copy*. Journal of Vegetation Science, 10(2):219–230, 1999.

[175] J. M. Ver Hoef. *Parametric empirical Bayes methods for ecological applications*. Ecological Applications, 6(4):1047–1055, 1996.

[176] K. Tucker, S. P. Rushton, R. A. Sanderson, E. B. Martin, and J. Blaiklock. *Modelling bird distributions — a combined GIS and Bayesian rule-based approach*. Landscape Ecology, 12(2):77–93, 1997.

[177] R. J. Aspinall. *An inductive modeling procedure based on Bayes theorem for analysis of pattern in spatial data*. International journal of geographical information systems, 6(2):105–121, 1992.

[178] H. S. Fisher. *Simulating the distribution of plant communities in an alpine landscape*. Coenoses, 5:37–43, 1990.

[179] G. M. Foody and M. E. J. Cutler. *Tree biodiversity in protected and logged Bornean tropical rain forests and its measurement by satellite remote sensing*. Journal of Biogeography, 30(7):1053–1066, 2003.

[180] S. S. Tan and F. E. Smeins. *Predicting grassland community changes with an artificial neural network model*. Ecological Modelling, 84(1–3):91–97, 1996.

[181] F. Recknagel, M. French, P. Harkonen, and K. Yabunaka. *Artificial neural network approach for modelling and prediction of algal blooms*. Ecological Modelling, 96(1–3):11–28, 1997.

[182] M. Hanewinkel, W. Zhoub, and C. Schilla. *A neural network approach to identify forest stands susceptible to wind damage*. Forest Ecology and Management, 196(2–3):227–243, 2004.

[183] G. M. Foody, P. M. Atkinson, P. W. Gething, N. A. Ravenhill, and C. K. Kelly. *Identification of specific tree species in ancient semi-natural woodland from digital aerial sensor imagery*. Ecological Applications, 15(4):1233–1244, 2005.

[184] G. M. Foody. *Applications of the self-organising feature map neural network in community data analysis*. Ecological Modelling, 120(2–3):97–107, 1999.

[185] S. L. Özemsi, C. O. Tan, and U. Özemsi. *Methodological issues in building, training, and testing artificial neural networks in ecological applications*. Ecological Modelling, 195(1–2):83–93, 2006.

[186] Q. H. Guo, M. Kelly, and C. H. Graham. *Support vector machines for predicting distribution of sudden oak death in California*. Ecological Modelling, 182(1):75–90, 2005.

[187] Y. Shan, D. Paull, and R. I. McKay. *Machine learning of poorly predictable ecological data*. Ecological Modelling, 195(1–2):129–138, 2006.

[188] J. M. Drake, C. Randin, and A. Guisan. *Modelling ecological niches with support vector machines*. Journal of Applied Ecology, 43(3):424–432, 2006.

[189] A. M. Prasad, L. R. Iverson, and A. Liaw. *Newer classification and regression tree techniques: bagging and random forests for ecological prediction*. Ecosystems, 9:181–199, 2006.

[190] Garzon M. B., R. Blazek, M. Neteler, R. S. de Dios, H. S. Ollero, and C. Furlanello. *Predicting habitat suitability with machine learning models: The potential area of Pinus silvestris L. in the Iberian Peninsula*. Ecological Modelling, 197(3–4):383–393, 2006.

[191] M. B. Araújo and M. News. *Ensemble forecasting of species distributions*. TRENDS in Ecology and Evolution, 22(1):42–47, 2007.

[192] R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. V. Tongeren, editors. *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge, 2nd edition, 1995.

[193] C. F. J. ter Braak. *Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis*. Ecology, 67:1167–1179, 1986.

[194] C. R. Rao. *The use and interpretation of principle component analysis in applied research*. Sankhya, A 26:329–358, 1964.

[195] A. L. van den Wollenberg. *Redundancy analysis. An alternative for canonical correlation analysis*. Psychometrika, 42:207–219, 1977.

[196] M. K-S. Tso. *Reduced-rank regression and canonical analysis*. Journal of the Royal Statistical Society Series B, 43:89–107, 1981.

[197] R. Gittins. *Canonical Analysis. A Review with Applications in Ecology*. Springer-Verslag, Berlin, 1985.

[198] R. H. Green. *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York, 1979.

[199] C. J. F. ter Braak. *Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model*. Biometrics, 41:859–873, 1985.

[200] D. Edwards. *Comment: the first data analysis should be journalistic*. Ecological Applications, 6(4):1090–1094, 1996.

[201] S. Lek and J. F. Guegan. *Artificial neural networks as a tool in ecological modelling, an introduction*. Ecological Modelling, 120(2–3):65–73, 1999.

[202] C. Cortes and V. Vapnik. *Support-vector networks*. Machine Learning, 20(3):273–297, 1995.

[203] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[204] L. Breiman. *Random forests*. Machine Learning, 45(1):5–32, 2001.

[205] J. R. Busby. *A biogeographical analysis of Nothofagus cunninghamii (Hook.) Oerst. in southern Australia*. Australian Journal of Ecology, 11:1–7, 1986.

[206] J. R. Busby. *BIOCLIM — A bioclimate analysis and prediction system*. In C. R. Margules and M. P. Austin, editors, Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, 1991.

[207] C. R. Cocks and I. A. Baird. *The role of Geographic Information Systems in the collection, extrapolation and use of survey data*. In C. R. Margules and M. P. Austin, editors, Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, 1991.

[208] G. Carpenter, A. N. Gillison, and J. Winter. *DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals*. Biodiversity Conservation, 2:667–680, 1993.

[209] J. Elith, C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. *Novel methods improve prediction of species' distributions from occurrence data*. Ecography, 29:129–151, 2006.

[210] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, Chichester, 2nd edition, 2000.

[211] V. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 2002.

[212] H. Akaike. *A new look at statistical model identification*. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[213] D. S. Moore. *Tests of chi-squared type*. In R. B. D'Agostino and M. A. Stevens, editors, Goodness-of-Fit Techniques. Marcel Dekker, New York, 1986.

[214] L. Hansen and P. Salamon. *Neural network ensembles*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12:993–1001, 1990.

[215] A. Liaw and M. Wiener. *Classification and regression by randomForest*. R News, 2–3:18–22, 2002.

[216] L. Breiman and A. Cutler. `http://www.stat./berkeley.edu/users/breiman/RandomForests/cc_software.htm`.

[217] P. Legendre. *Spatial autocorrelation: trouble or new paradigm?* Ecology, 74(6):1659–1673, 1993.

[218] L. Breiman and A. Cutler. `http://www.stat./berkeley.edu/users/breiman/RandomForests/cc_papers.htm`.

[219] I. P. Vaughan and S. J. Ormerod. *The continuing challenges of testing species distribution models*. Journal of Applied Ecology, 42:720–730, 2005.

[220] J. Cohen. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20:37–46, 1960.

[221] T. G. Dietterich. *Approximate statistical tests for comparing supervised classification learning algorithms*. Neural Computation, 10:1895–1923, 1998.

[222] B. S. Everitt. *The analysis of contingency tables*. Chapman and Hall, London, 2nd edition, 1992.

[223] A. H. Fielding and J. F. Bell. *A review of methods for the assessment of prediction errors in conservation presence/absence models*. Environmental Conservation, 24(1):38–49, 1997.

[224] R. Kohavi and F. Provost. *Glossary and terms. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process.* Machine Learning, 30(2–3):271–274, 1998.

[225] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[226] F. Wilcoxon. *Individual comparisons by ranking methods*. Biometrics, 1:80–83, 1945.

[227] T. Fawcett. *An introduction to ROC analysis*. Pattern Recognition Letters, 27:861–874, 2006.

[228] R. G. Pontius Jr. and L. Schneider. *Land-use change model validation by a ROC method for the Ipswich watershed, Massachusetts, USA*. Agriculture, Ecosystems & Environment, 85(1–3):239–248, 2001.

[229] M. S. Boyce, P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. *Evaluating resource selection functions*. Ecological modelling, 157(2–3):281–300, 2002.

[230] C. R. Liu, P. M. Berrey, T. P. Dawson, and R. G. Pearson. *Selecting thresholds of occurrence in the prediction of species distributions*. Ecography, 28(3):385–393, 2005.

[231] S. J. Phillips, R. P. Anderson, and R. E. Schapire. *Maximum entropy modeling of species geographic distributions*. Ecological modelling, 190(3–4):231–259, 2006.

[232] R. Díaz-Uriarte and S. Alvarez de Andrés. *Gene selection and classification of microarray data using random forest*. BMC Bioinformatics, 7(3):doi:10.1186/1471–2105–7–3, 2006.

[233] X. W. Chen and M. Liu. *Prediction of protein-protein interactions using random decision forest framework*. Bioinformatics, 21(24):4394–4400, 2006.

[234] M. Pal. *Random forest classifier for remote sensing classification*. International Journal of Remote Sensing, 26(1):217–222, 2005.

[235] J. Ham, Y. C. Chen, M. P. Crawford, and J. Ghosh. *Investigation of the random forest framework for classification of hyperspectral data*. IEEE Transactions on Geoscience and Remote Sensing, 43(3):492–501, 2005.

[236] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. *Random Forests for land cover classification*. Pattern Recognition letters, 27(4):294–300, 2006.

[237] A. Guisan and W. Thuiller. *Predicting species distribution: offering more than simple habitat models*. Ecology Letters, 8:993–1009, 2005.

[238] R. Mac Nally. *Regression and model-building in conservation biology, biogeography, and ecology: the distinction between – and reconciliation of – 'predictive' and 'explanatory' models*. Biodiversity and Conservation, 9:655–671, 2000.

[239] M. G. Turner, S. E. Gergel, M. D. Dixon, and J. R. Miller. *Distribution and abundance of trees in floodplain forests of the Wisconsin river: environmental influences at different scales*. Journal of Vegetation Science, 15:729–738, 2004.

[240] I. N. Vogiatzakis, G. H. Griffiths, and A. M. Mannion. *Environmental factors and vegetation composition, Lefka Ori massif, Crete, S. Aegean*. Global Ecology and Biogeography, 12:131–146, 2003.

[241] J. Lyon and N. M. Gross. *Patterns of plant diversity and plant-environment relationships across three riparian corridors*. Forest Ecology and Management, 204:267–278, 2005.

[242] R. Mac Nally. *Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables*. Biodiversity and Conservation, 11:1397–1401, 2002.

[243] A. Guisan, J. Edwards, C. Thomas, and T. Hastie. *Generalized linear and generalized additive models in studies of species distributions: setting the scene*. Ecological Modelling, 157:89–100, 2002.

[244] A. Chevan and M. Sutherland. *Hierarchical partitioning*. The American Statistician, 45(2):90–96, 1991.

[245] E. Fleisman, R. Mac Nally, and D. D. Murphy. *Relationships among non-native plants, diversity of plants and butterflies, and adequacy of spatial sampling*. Biological Journal of the Linnean Society, 85:157–166, 2005.

[246] M. O. Hill and H. G. Gaugh. *Detrended correspondence analysis, an improved ordination technique*. Vegetatio, 42:47–58, 1980.

[247] R. H. Whittaker. *Gradient analysis of vegetation*. Biological Reviews, 49:207–264, 1967.

[248] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. WCB McGraw-Hill, United States, 4th edition, 1996.

[249] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2002.

[250] C. Walsh and R. Mac Nally. `http://www.cran.r-project.org/doc/packages/hier.part.pdf`.

[251] P. Legendre and L. Legendre. *Numerical ecology*. Elsevier Science, Amsterdam, 2nd edition, 1998.

[252] C. Strobl, A-L. Boulesteix, A. Zeileis, and T. Hothorn. *Bias in random forest variable importance measures: Illustrations, sources and a solution*. BMC Bioinformatics, 8(25):doi:10.1186/1471–2105–8–25, 2007.

[253] K. J. Archer and R. V. Kimes. *Empirical characterization of random forest variable importance measures*. Computational Statistics & Data Analysis, 52:2249–2260, 2008.

[254] M. J. Wassen and A. Barendregt. *Topgraphic position and water chemistry of fens in a Dutch riverplain*. Journal of Vegetation Science, 3:447–456, 1992.

[255] N. S. Philippi D. L. Hey. *Flood reduction through wetland restoration: The upper Mississippi river basin as a case history*. Restoration Ecology, 3:4–17, 1995.

[256] O. L. Gilbert and P. Anderson. *Habitat Creation and Repair*. Oxford University Press, Oxford, 1998.

[257] R. J. Hobbs and D. A. Norton. *Towards a conceptual framework for restoration ecology*. Restoration Ecology, 4:93–110, 1996.

[258] H. O. Venterink, M. J. Wassen, J. D. M. Belgers, and J. T. A. Verhoeven. *Control of environmental variables on species density in fens and meadows: importance of direct effects and effects through community biomass*. Journal of Ecology, 89(6):1033–1040, 2001.

[259] M. J. Wassen, W. H. M. Peeters, and H. O. Venterink. *Patterns in vegetation, hydrology, and nutrient availability in an undisturbed river floodplain in Poland*. Plant Ecology, 165(1):27–43, 2005.

[260] A. R. Hill. *Nitrate removal in stream riparian zones*. Journal of Environmental Quality, 25(4):743–755, 1996.

[261] J. Fisher and M. C. Acreman. *Wetland nutrient removal: a review of the evidence*. Hydrology and Earth System Sciences, 8(4):673–685, 2006.

[262] P. Jaccard. *The distribution of the flora of the alpine zone*. New Phytologist, 11:37–50, 1912.

[263] W. Huybrechts, P. De Becker, E. De Bie, E. Wassen, and A. Bio. *Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, ITORS-VL. VLINA 00/16*. Institute of Nature Conservation, Brussels, Belgium, 2002. (In Dutch).

[264] F. Provost and P. Domingos. *Well-trained PETs: Improving probability estimation trees*. CeDER Working Paper #IS-00-04, 2001.

[265] S. Ferrier. *Mapping spatial pattern in biodiversity for regional conservation planning: where to from here?* Systematic Biology, 51:331–363, 2002.

[266] S. Barry and J. Elith. *Error and uncertainty in habitat models*. Journal of Applied Ecology, 43:413–423, 2006.

[267] D. L. Phillips and D. G. Marks. *Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models*. Ecological Modelling, 91(1–3):213–229, 1996.

[268] P. W. van Horssen, E. J. Pebesma, and P. P. Schot. *Uncertainties in spatially aggregated predictions from a logistic regression model*. Ecological Modelling, 154(1–2):93–101, 2002.

[269] T. Larssen, T. Høgåsen, and B. J. Cosby. *Impact of time series data on calibration and prediction uncertainty for a deterministic hydrogeochemical model*. Ecological Modelling, 207(1):22–33, 2007.

[270] K. P. Van Niel and M. P. Austin. *Predictive vegetation modelling for conservation: Impact of error propagation from digital elevation data*. Ecological Applications, 17(1):266–280, 2007.

[271] N. Ray and M. A. Burgman. *Subjective uncertainties in habitat suitability maps*. Ecological Modelling, 195(3–4):172–186, 2006.

[272] S. Dray, P. Legendre, and P. R. Peres-Neto. *Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM)*. Ecological Modelling, 196(3–4):483–493, 2006.

[273] D. Borcard, P. Legendre, and P. Drapeau. *Partialling out the spatial component of ecological variation*. Ecology, 73(3):1045–1055, 1992.

[274] D. Borcard and P. Legendre. *All-scale analysis of ecological data by means of principal coordinates of neighbour matrices*. Ecological Modelling, 153(1–2):51–68, 2002.

[275] D. Borcard, P. Legendre, C. Avois-Jacquet, and H. Tuomisto. *Dissecting the spatial structure of ecological data at multiple scales*. Ecology, 85(7):1826–1832, 2004.

[276] J. C. Gower. *Some distance properties of latent root and vector methods used in multivariate analysis*. Biometrika, 53:325–338, 1966.

[277] G. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, UK, 1997.

[278] M. Alfaro. *Étude de la robustesse des simulations de fonctions aléatoires*. PhD thesis, E.N.S. des Mines de Paris, 1979.

[279] H. Bourennane, D. King, A. Couturier, B. Nicoullaud, B. Mary, and G. Richard. *Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: An empirical comparison of simulation accounting for single attribute and a simulation accounting for secondary information*. Ecological Modelling, 205(3–4):323–335, 2007.

[280] M. Fagrout and M. Van Meirvenne. *Accounting for soil autocorrelation in the design of experimental trials*. Soil Science Society of America Journal, 66:1134–1142, 2002.

[281] C. Deutsch and A. Journel. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, UK, 2nd edition, 1998.

[282] G. Goovaerts. *Geostatistical modeling of uncertainty in soil science*. Geoderma, 103:3–26, 2001.

[283] H. G. Gauch and R. H. Whittaker. *Hierarchical classification of community data*. Journal of Ecology, 69:537–557, 1981.

[284] C. E. Shannon. *A mathematical theory of communication*. Bell System Technical Journal, 27:379–423, 1948.

[285] C. E. Shannon. *A mathematical theory of communication*. Bell System Technical Journal, 27:623–656, 1948.

[286] E. Van Broekhoven, V. Adriaenssens, B. De Baets, and P. F. M. Verdonschot. *Fuzzy rule-based macroinvertebrate habitat suitability models for running waters*. Ecological Modelling, 198(1–2):71–84, 2006.

[287] J. Oksanen and P. R. Minchin. *Spatial complexity of ecological communities: Bridging the gap between probabilistic and non-probabilistic uncertainty measures*. Ecological Modelling, 197(1–2):59–66, 2006.

[288] J. Miller, J. Franklin, and R. Aspinall. *Incorporating spatial dependence in predictive vegetation models*. Ecological Modelling, 202(3–4):225–242, 2007.

[289] M. D. McKay, R. J. Beckman, and W. J. Conover. *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*. Technometrics, 21:239–245, 1979.

[290] R. L. Imam and W. J. Conover. *Small sample sensitivity analysis techniques for computer models, with an application to risk assessment*. Communications in Statistics Theory and Methods, A9:1749–1874, 1980.

[291] B. Minasny and A. B. McBradney. *A conditioned Latin hypercube method for sampling in the presence of ancillary information*. Computers & Geosciences, 32:1378–1388, 2006.

[292] C. J. F. ter Braak, H. Hoijtink, W. Akkermans, and P. F. M. Verdonschot. *Bayesian model-based cluster analysis for predicting macrofaunal communities*. Ecological Modelling, 160(3):235–248, 2003.

[293] H. Vangroenewoud. *The robustness of correspondence, detrended correspondence, and TWINSPAN analysis*. Journal of Vegetation Science, 3(2):239–246, 1992.

[294] J. Oksanen and P. R. Minchin. *Instability of ordination results under changes in input data order: Explanations and remedies*. Journal of Vegetation Science, 8(3):447–454, 1997.

[295] K. Kleinod, M. Wissen, and M. Bock. *Vegetation mapping and nature conservation in Switzerland*. Plant Ecology, 110(1):19–23, 1994.

[296] B. Wood. *Room for nature? Conservation management of the Isle of Rum, UK and prospects for large protected areas in Europe*. Biological Conservation, 94(1):93–105, 2000.

[297] E. Dias, R. B. Elias, and V. Nunes. *Vegetation mapping and nature conservation: a case study in Terceira Island (Azores)*. Biodiversity and Conservation, 13(8):1519–1539, 2004.

[298] K. Kleinod, M. Wissen, and M. Bock. *Detecting vegetation changes in a wetland area in Northern Germany using earth observation and geodata*. Journal of Nature Conservation, 13:115–125, 2005.

[299] P. Pérez Ballesta. *The uncertainty of averaging a time series of measurements and its use in environmental legislation*. Atmospheric Environment, 39:2003–2009, 2005.

[300] P. L. Bjerg and T. H. Christensen. *Spatial and temporal small-scale variation in groundwater quality of a shallow sandy aquifer*. Journal of Hydrology, 131(1–4):133–149, 1992.

[301] I. Varsanyi. *Temporal variability in groundwater chemistry in the great Hungarian plain during the period 1975–1989*. Hydrological Sciences Journal - Journal des Sciences Hydrologiques, 37(2):119–128, 1992.

[302] D. Boeye, L. Clement, and R. F. Verheyen. *Hydrochemical variation in a groundwater discharge fen*. Wetlands, 14(2):122–133, 1994.

[303] A. E. Williams, J. A. Johnson, L. J. Lund, and Z. J. Kabala. *Spatial and temporal variations in nitrate contamination of a rural aquifer, California*. Journal of Environmental Quality, 27(5):1147–1157, 1998.

[304] J. Gelbrecht, H. Lengsfeld, R. Pothig, and D. Opitz. *Temporal and spatial variation of phosphorus input, retention an loss in a small catchment of NE Germany*. Journal of Hydrology, 304(1–4):151–165, 2005.

[305] Wetlands International. `http://www.wetlands.org`.

[306] P. Dugan. *Wetlands in Danger*. Michael Beasly, Reed International Books, London, 1993.

[307] Convention on Wetlands of International Importance. `http://www.ramsar.org`.

[308] W. M. Childress, C. M. Crisafulli, and E. J. Jr. Rykiel. *Comparison of Markovian matrix models of a primary successional plant community*. Ecological Modelling, 107:93–102, 1998.

[309] V. N. Korotkov, D. O. Logofet, and M. Loreau. *Succession in mixed boreal forest of Russia: Markov models and non-Markov effects*. Ecological Modelling, 142:25–38, 2001.

[310] L. Orlóci and M. Orlóci. *On recovery, Markov chains, end canonical analysis*. Ecology, 69(4):1260–1265, 1988.

[311] M. Usher. *Modelling ecological succession, with particular reference to Markovian models*. Vegetatio, 46:11–18, 1981.

[312] H. Balzter, P. Braun, and W. Köhler. *Cellular automate models for vegetation dynamics*. Ecological Modelling, 107:113–125, 1998.

[313] B. Tucker and M. Anand. *The use of matrix models to detect natural and pollution-induced forest gradients*. Community Ecology, 4(1):89–110, 2003.

[314] E. Lippe, J. De Smidt, and D. Glenn-Lewin. *Markov models and succession: a test from a heathland in the Netherlands*. Journal of Egology, 73:775–791, 1985.

[315] D. Hartfiel. *Component bounds on Markov set-chain limiting sets*. Journal of Statistical Computation Simulation, 38:15–24, 1991.

[316] L. Baum and T. Petrie. *Statistical inference for probabilistic functions of finite state Markov chains*. Bulletin of the American Mathematical Society, 73:1554–1563, 1966.

[317] O. Batelaan, F. De Smedt, and L. Triest. *Regional groundwater discharge: phreatophyte mapping, groundwater modelling and impact analysis of land-use change*. Journal of Hydrology, 275:86–108, 2003.

[318] O. Batelaan and F. De Smedt. *WetSpass: a flexible, GIS based, distributed recharge methodology for regional groundwater modelling*. In H. Gehrels, J. Peters, E. Hoehn, K. Jensen, C. Leibundgut, J. Giffioen, B. Webb, and W. J. Zaadnoordijk, editors, Impact of Human Activity on Groundwater Dynamics, number 269 in IAHS. Wallingford, 2001.

[319] O. Batelaan and F. De Smedt. *An adapted DRAIN package for seepage problems*. In E. Poeter, C. Zheng, and M. Hill, editors, MODFLOW'98 Proceedings, volume 2, Golden, 1998. Colorado School of Mines.

[320] M. G. McDonald and A. W. Harbaugh. *A modular three-dimensional finite-difference ground-water flow model*, volume Techniques of Water-Resources Investigations. US Geological Survey, Reston, Virginia, 1998.

[321] J. Peters, R. Samson, P. Boeckx, W. K. Balasooriya, and N. E. C. Verhoest. *Hydrologic and vegetational analysis of an alluvial floodplain in Belgium*. Belgian Journal of Botany, 2007. Submitted.

[322] N. Breda and A. Granier. *Intra- and interannual variations of transpiration, leaf area index and radial growth of a sessile oak stand (Quercus petraea)*. Annales des Sciences Forestieres, 53(2–3):521–536, 1996.

[323] S. Shimoda and T. Oikawa. *Characteristics of canopy evapotranspiration from a small heterogeneous grassland using thermal imaging*. Environmental and Experimental Botany, 63(1–3):102–112, 2008.

[324] D. M. Sumner and J. M. Jacobs. *Utility of Penman-Monteith, Priestley-Taylor, reference evapotranspiration, and pan evaporation methods to estimate pasture evapotranspiration*. Journal of Hydrology, 308(1–4):81–104, 2005.

[325] A. Tuzet, J. F. Castell, A. Perrier, and O. Zurfluh. *Flux heterogeneity and evapotranspiration partitioning in a sparse canopy: The fallow savanna*. Journal of Hydrology, 189(1–4):482–493, 1997.

[326] D. D. Baldocchi, B. E. Law, and P. M. Anthoni. *On measuring and modeling energy fluxes above the floor of a homogeneous and heterogeneous conifer forest*. Agricultural and Forest Meteorology, 102(2–3):187–206, 2000.

[327] A. Zwaenepoel, F. T'jollyn, V. Vandenbussche, and M. Hoffman. *Systematiek van natuurtypen voor Vlaanderen: 6.3 graslanden, natte hooilanden op (matig) voedselarme gronden*. Instituut voor Natuurbehoud, Brussel, 2002. (In Dutch).

A

Shortlist of plant species

PHREATOPHYTES

Achillea ptarmica L.
Carex canescens L.
Carex nigra (L.) Reichard
Equisetum telmateia Ehrh.
Filipendula ulmaria (L.) Maxim.
Lythrum salicaria L.
Myrica gale L.
Osmunda regalis L.
Ribes nigrum L.
Scirpus sylvaticus L.
Scutellaria galericulata L.
Thalictrum flavum L.
Angelica sylvestris L.
Calamagrostis canescens (Weber) Roth
Circaea lutetiana L.
Dactylorhiza fistulosa (Moench) H. Baumann et Künkele
Deschampsia cespitosa (L.) Beauv.
Lotus pedunculatus Cav.
Luzula multiflora (Ehrh.) Lej.

Adapted from W. Huybrechts, E. De Bie, P. De Becker, M. Wassen, and A. Bio. *Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, ITORS-VL (VLINA 00/16).* Instituut voor Natuurbehoud, Brussel, 2002. (In Dutch)

Lysimachia vulgaris L.
Phalaris arundinacea L.
Pulicaria dysenterica (L.) Bernh.
Saxifraga granulata L.
Alopecurus pratensis L.
Barbarea intermedia Boreau
Carex sylvatica Huds.
Lamium galeobdolon (L.) L.
Molinia caerulea (L.) Moench
Rhinanthus angustifolius C.C. Gmel.
Rhinanthus minor L
Alisma plantago-aquatica L.
Berula erecta (Huds.) Coville
Calla palustris L.
Caltha palustris L.
Carex acuta L.
Carex acutiformis Ehrhr.
Carex disticha Huds.
Carex echinata Murray
Carex elongata L.
Carex paniculata L.
Carex pseudocyperus L.
Carex rostrata Stokes
Cirsium oleraceum (L.) Scop.
Comarum palustre L.
Eleocharis palustris (L.) Roem. et Schult.
Equisetum fluviatile L.
Equisetum palustre L.
Eriophorum polystachion L.
Galium palustre L.
Galium uliginosum L.
Glyceria maxima (Hartm.) Holmberg
Juncus acutiflorus Ehrh. ex Hoffmann
Juncus filiformis L.
Menyanthes trifoliata L.
Peucedanum palustre (L.) Moench
Phragmites australis (Cav.) Steud.
Ranunculus flammula L.
Rumex hydrolapathum Huds.
Viola palustris L.

NON-PHREATOPHYTES

Anthriscus sylvestris (L.) Hoffmann
Arrhenatherum elatius (L.) Beauv. Ex J. et C. Presl
Brachypodium sylvaticum Boreau
Carex hirta L.
Crepis biennis L.
Dactylorhiza maculata (L.) Soó
Deschampsia flexuosa (L.) Trin.
Digitalis purpurea L.
Equisetum arvense L.
Geum urbanum L.
Oxalis fontana Bunge
Polygonatum multiflorum (L.) All.
Potentilla sterilis (L.) Garcke
Pteridium aquilinum (L.) Kuhn
Teucrium scorodonia L.
Vaccinium myrtillus L.

B

Photograph of the different vegetation types

ALNO - PADION, *Elzen - Vogelkers verbond*, [elzen - vogelkers bos]

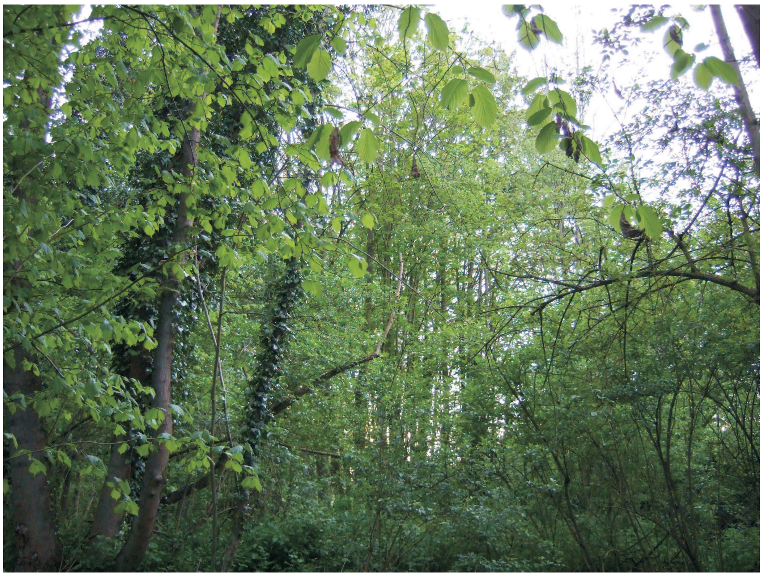


FIGURE B.1 – The forest type *Alno - Padion* is present at Snoekengracht (April 2008).

ARRHENATHERION ELATIORIS, *Glanshaver verbond*, [glanshaverhooiland]

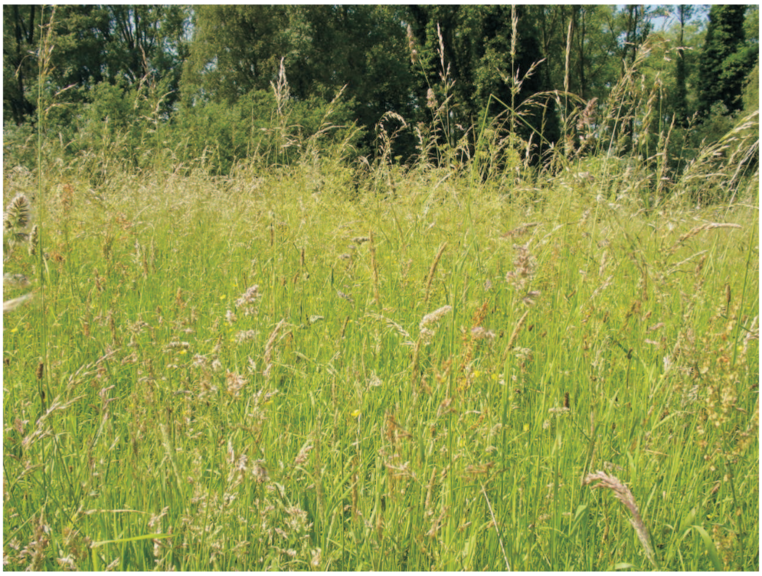


FIGURE B.2 – *Arrhenatherion elatioris* with several grass species, as observed at Bourgoyen-Ossemeersen, Gent.

CALTHION PALUSTRIS , *Dotterbloem verbond*, [dottergrasland]

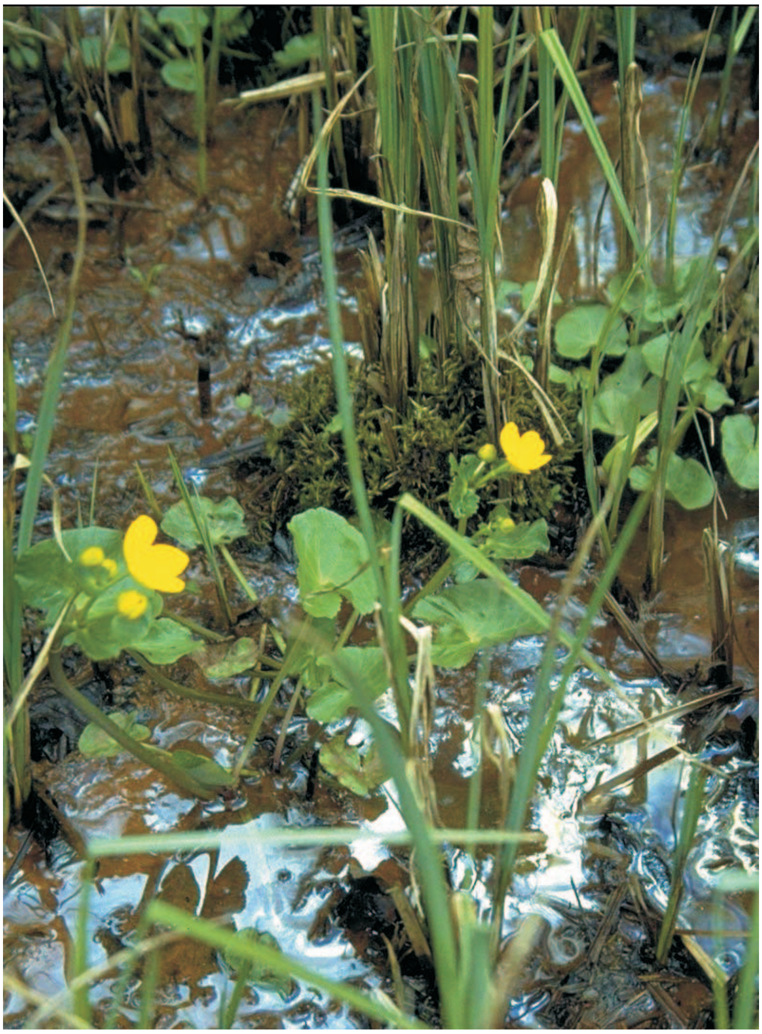


FIGURE B.3 – *Caltha palustris* is a diagnostic species for the *Calthion palustris* vegetation type. The blueish aspect of the water on the soil surface is characteristic for seepage water. (Copied with permission from [?]).

CARICI ELONGATAE – ALNETUM GLUTINOSAE, *Elzenzegge - Elzenbroek associatie*, [mesotroof elzenbroek]

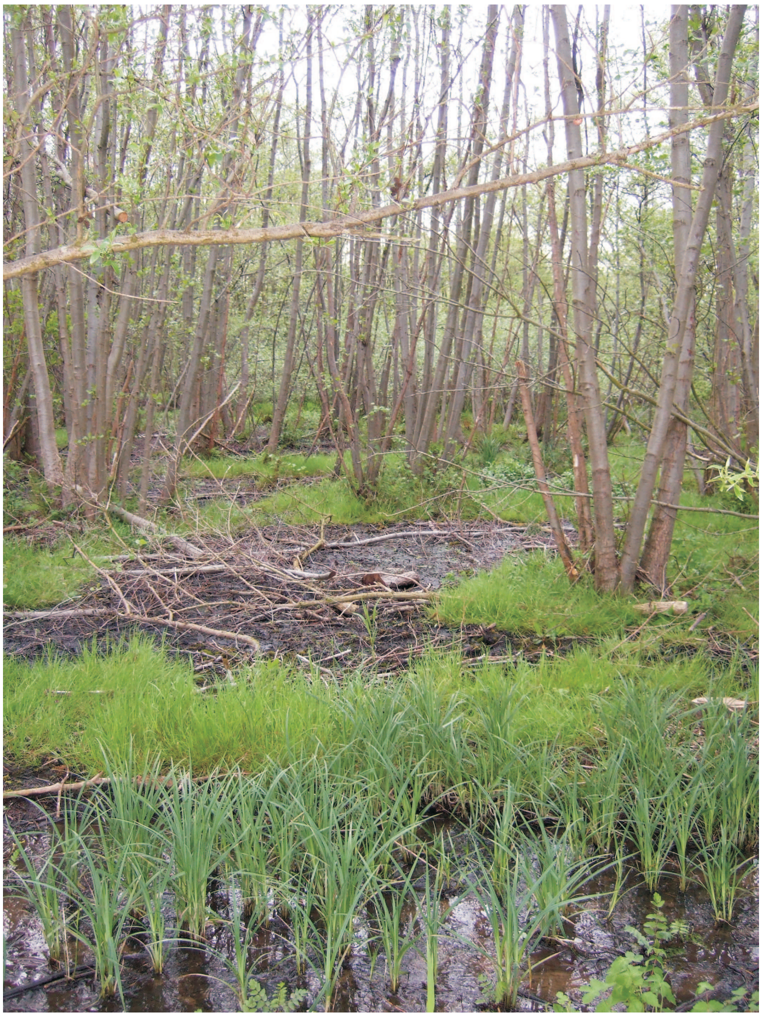


FIGURE B.4 – *Carici elongatae – Alnetum glutinosae* at Snoekengracht (April 2008). An *Alnus glutinosa* tree layer is undergrown by a herblayer with *Carex acutiformis*, as can be seen on the foreground of this photograph.

CARICION CURTO-NIGRAE, *Zomp -en zwarte zegge verbond*, [kleine zeggevegetatie]

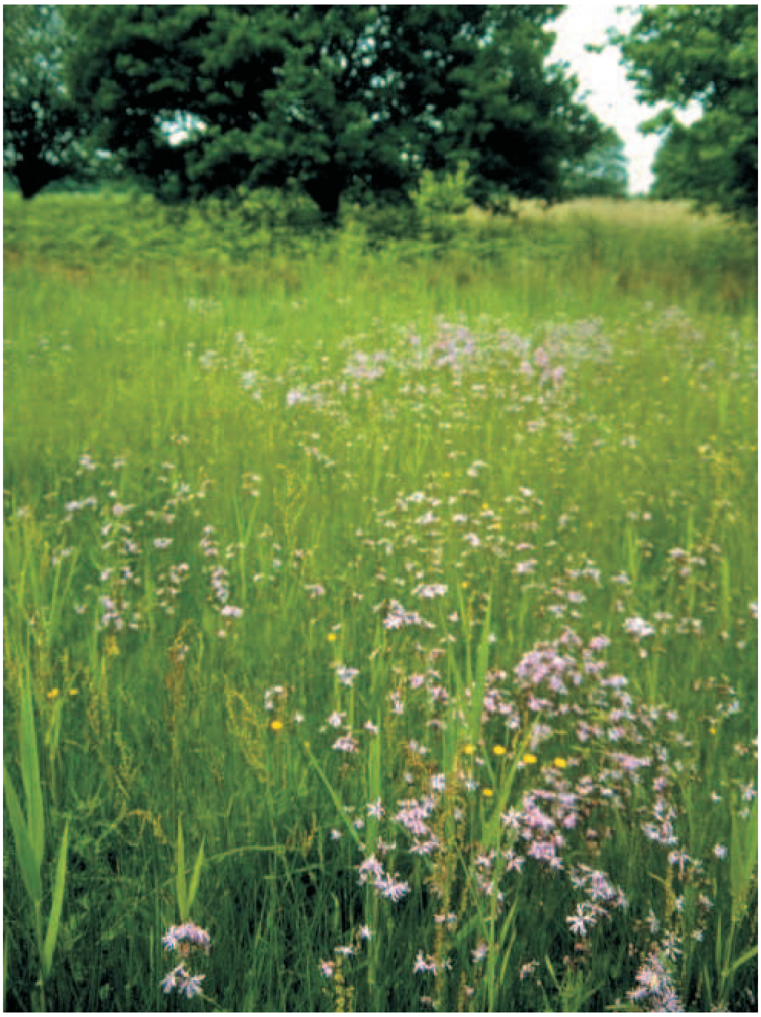


FIGURE B.5 – *Caricion curto-nigrae* with flowering *Lychnis flos-cuculi* at Zwarte Beek has a high species similarity with *Calthion palustris*. (Copied with permission from [?]).

CIRSIO – MOLINIETUM , *Associatie van Kale Jonker en Pijpestrootje*, [blauwgrasland]

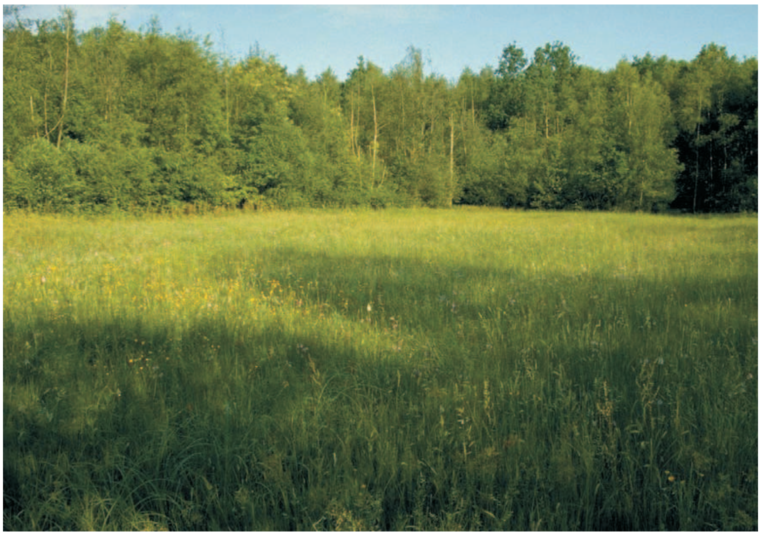


FIGURE B.6 – Two adjacent vegetation types at Vorsdonkbos-Turfputten: the herbaceous *Cirsio – Molinietum* on the foreground and *Sphagno – Alnetum*, the woody vegetation type on the background. (Copied with permission from [?]).

FILIPENDULION, *Moerasspirrea verbond*, [moerasspirearuigte]

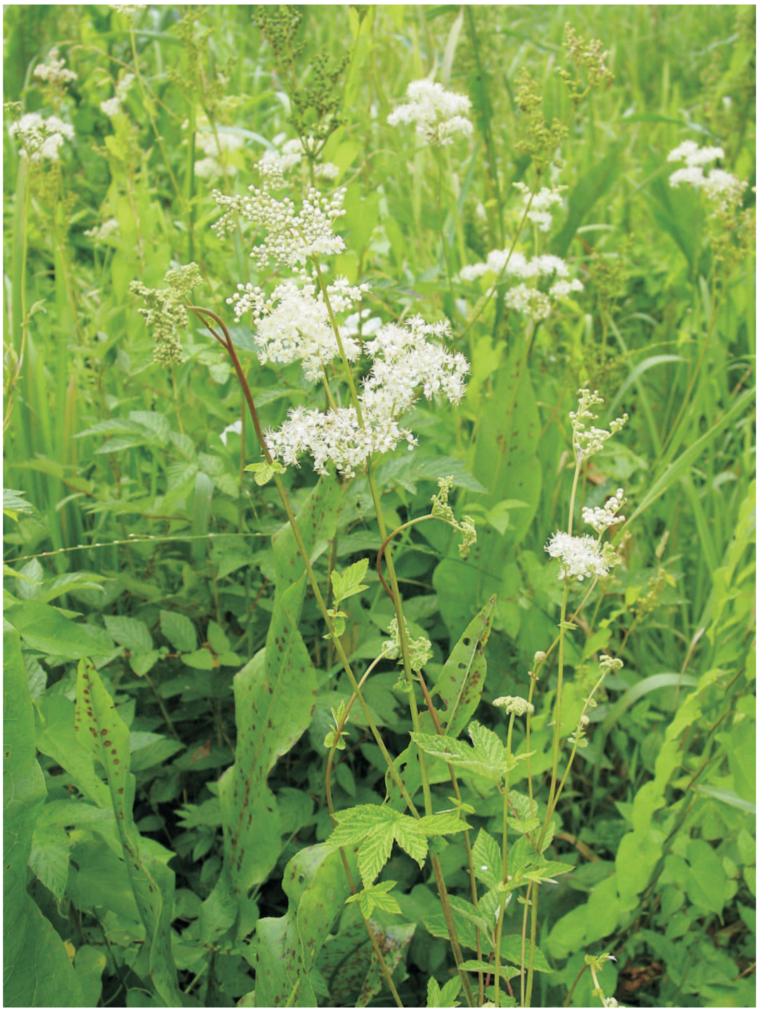


FIGURE B.7 – *Filipendula ulmaria* is a diagnostic species for the *Filipendulion* vegetation type.

MAGNOCARICION, *verbond van de grote zegge soorten*, [grote zeggevegetatie]

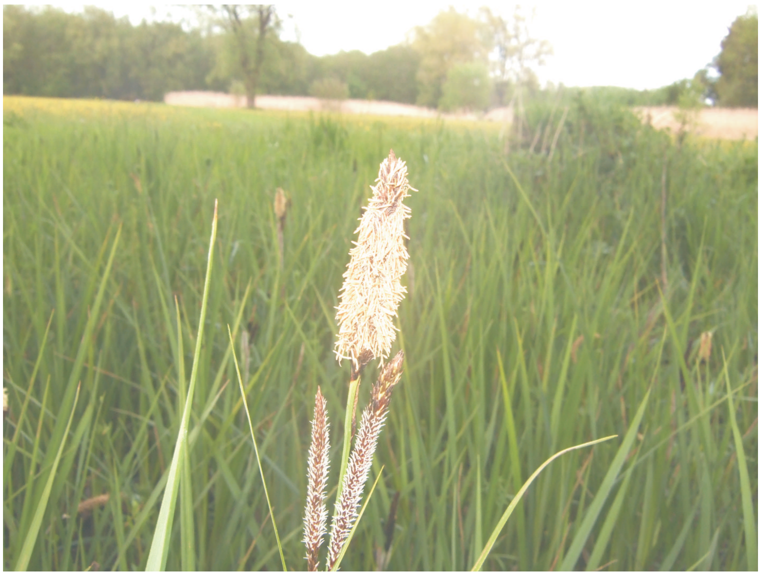


FIGURE B.8 – *Magnocaricion* vegetation type, with detail of a flowering *Carex* species.

MAGNOCARICION WITH PHRAGMITES , *verbond van de grote zegge soorten met riet*, [rietruigte]

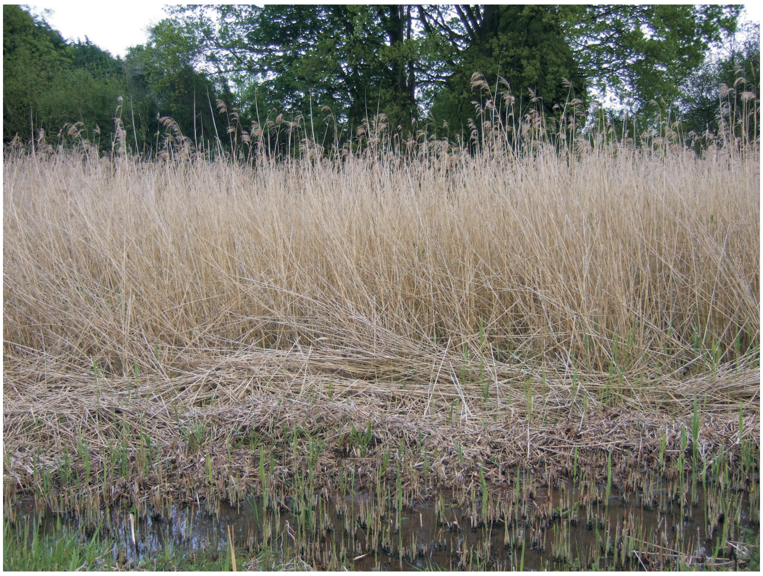


FIGURE B.9 – *Magnocaricion* with *Phragmites* at Snoekengracht (foreground, April 2008).

PHRAGMITETALIA, *Riet orde*, [rietland]

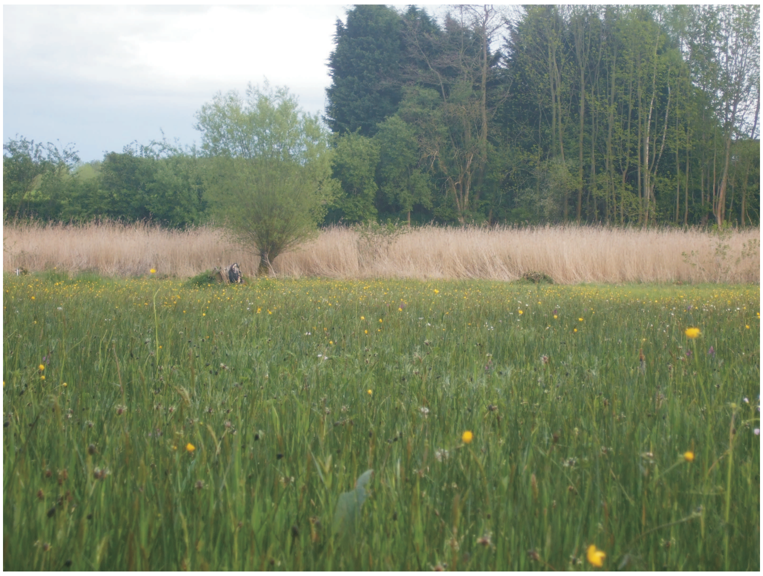


FIGURE B.10 – The brown *Phragmitetalia* vegetation belt (winter aspect) at Snoekengracht (April 2008).

SPHAGNO – ALNETUM, [oligotroof elzen - berkenbos]
(See Fig. B.6)

Jan Peters

Curriculum Vitae

Contact

Lentestraat 22
9000 Gent
Belgium

Tel.: +32 486 299 112

Jan.Peters@Ugent.be

Opleiding

Bio-ingenieur, juni 2002
Katholieke Universiteit Leuven
Leuven, België

Thesis: *Comparison of different Life Cycle Analysis methods for land use impact assessment in a Pinus radiata plantation (Jonkershoek, South Africa)*
Promotor: Prof. Bart Muys

Tewerkstelling

10/2004–10/2008 Wetenschappelijk medewerker
Laboratorium voor Hydrologie en Waterbeheer
Vakgroep Bos- en Waterbeheer
Universiteit Gent
Gent, België

3/2003–9/2004 Wetenschappelijk medewerker
Peat Technology Centre
Biosystems Engineering Department University College Dublin
Dublin, Ireland

Wetenschappelijk Curriculum

Internationale publicaties met peer review

W. K. Balasooriya, K. Denef, J. Peters, N. E. C. Verhoest, P. Boeckx. *Vegetation composition and soil microbial community structural changes along a wetland hydrological gradient*. Hydrology and Earth System Science, 12:277–291, 2008.

J. Peters, B. De Baets, R. Samson and N. E. C. Verhoest. *Modelling groundwater-dependent vegetation patterns using ensemble learning.* Hydrology and Earth System Sciences, 12:603–613, 2008.

J. Peters, B. De Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. De Becker and W. Huybrechts. *Random forests as a tool for ecohydrological distribution modelling.* Ecological Modelling, 207:304–318, 2007.

J. F. García-Quijano, J. Peters, L. Cockx, G. van Wyk, A. Rosanov, G. Deckmyn, R. Ceulemans, S. M. Ward, N. M. Holden, J. Van Orshoven, and B. Muys. *Carbon sequestration and environmental effects of afforestation with Pinus radiata D. Don in the Western Cape, South Africa*. Climatic Change, 83:323–355, 2007.

J. Peters, V. Wieme, P. Boeckx, R. Samson, R. Godoy, C. Oyarzun, and N. Verhoest. *Possibilities for ecohydrological monitoring in natural and managed ecosystems in Southern Chile*. Gayana Botanica, 62(2):120–129, 2005.

J. Peters, N. E. C. Verhoest, B. De Baets, R. Samson and P. Boeckx. *Wetland vegetation distribution modelling for the identification of constraining environmental variables.* Landscape Ecology, accepted, 2008 (DOI : 10.1007/s10980-008-9261-4).

J. Peters, N. E. C. Verhoest, R. Samson, M. Van Meirvenne, L. Cockx, Z. Vekerdy and B. De Baets. *Uncertainty propagation in vegetation distribution models based on ensemble learning.* Ecological Modelling, submitted, 2008.

J. Peters, R. Samson, P. Boeckx, W. K. Balasooriya, and N. E. C. Verhoest. *Hydrologic and vegetational analysis of an alluvial floodplain in Belgium*. Belgian Journal of Botany, submitted, 2007.

Nationale publicaties met peer review

J. Peters, N. E. C. Verhoest, B. De Baets, P. De Becker, W. Huybrechts, and R. Samson. *Regressie- en classificatietechnieken in hydro-ecologische modellen: toepassing voor vallei-ecosystemen in Vlaanderen.* Water 31–'Bodem, Grondwater en Ecosysteem', 2007.

J. Peters, R. Samson, and N. E. C. Verhoest. *Predictive ecohydrological modelling using the random forest algorithm*. Communications in Agricultural and Applied Biological Science, 70(2):207–211, 2005.

Bijdrage op internationale conferenties en workshops

J. Peters, J. García Quijano, T. Content, G. Van Wyk, N. M. Holden, S. M. Ward, and B. Muys. *A new land use impact assessment method for LCA: theoretical fundaments and field validation*. 4th International Conference on life cycle assessment in the agri-food sector. Linking environmentally friendly production and sustainable consumption. Denmark, October 6-8, 2003, p. 138. (oral presentation)

J. Peters, B. Olivie, N. M. Holden, S. M. Ward, and B. Muys. *Application of the concept of exergy in land use impact assessment*. Faculty of agri-food and the environment research report 2002-2003, p. 154-157.

J. Peters, S. M. Ward, and N. M. Holden. *A novel life cycle impact assessment (LCIA) methodology for land use on Irish peatlands*. 14th Irish Environmental Researchers' Colloquium, Environ, 2004, p. 22. (oral presentation)

J. Peters, S. M. Ward, and N. M. Holden. *A novel life cycle impact assessment (LCIA) methodology for land use on Irish peatlands*. Proceedings of the 12th International Peat Congress: Wise Use of Peatlands. Finland, June 6–11, 2004, p. 478–483. (oral presentation)

J. Peters, V. Wieme, N. E. C. Verhoest, B. De Baets, R. Samson, P. Boeckx, P. De Becker, and W. Huybrechts. *Random forests as a tool for predictive ecohysrological modelling*. International Symposium on Wetland Pollutant Dynamics and Control, Belgium, September 4–8, 2005, p. 137–138. (oral presentation)

R. Samson, C. Van Gucht, J. Peters, N. E. C. Verhoest, V. Wieme, G. Spanoghe, F. M. G. Tack, and P. Boeckx. *Contolled inundation in the wetland 'De Assels' (East-Flanders, Belgium)? An ecohydrological assessment*. International Symposium on Wetland Pollutant Dynamics and Control, Belgium, September 4–8, 2005, p. 283–284. (poster presentation)

V. Wieme, J. Peters, N. E. C. Verhoest, R. Samson, and P. Boeckx, P. *Ecohdrological monitoring of a transect in the wetland 'Bourgoyen-Ossemeersen' (Ghent, Belgium)*. International Symposium on Wetland Pollutant Dynamics and Control, Belgium, September 4–8, 2005, p. 294–295. (poster presentation)

J. Peters, N. E. C. Verhoest, B. De Baets, R. Samson, and P. Boeckx. *'Random forests' as a distribution modelling technique in ecohydrology*. HydroEco'2006, International Conference on Hydrology and Ecology: The Groundwater/Ecology Connection. Karlovy Vary, Czech Republic, 11-14 September, 2006. (Poster presentation)

J. Peters, N. E. C. Verhoest, R. Samson, B. De Baets, and P. Boeckx. *On the identification of important predictors in ecohydrological modelling*. HydroEco'2006: International Conference on Hydrology and Ecology: The Groundwater–Ecology Connection. Karlovy Vary, Czech Republic, 11-14 September, 2006. (Oral presentation)

J. Peters, N. E. C. Verhoest, B. De Baets, and R. Samson. *The random forests technique: an application in eco-hydrologic distribution modeling*. European Geosciences Union, General Assembly 2007, Vienna, Austria, April 15-20, 2007. (Poster presentation in session HS46: Hydroinformatics: computational intelligence and technological developments in water science applications)

J. Peters, N. E. C. Verhoest, R. Samson, and P. Boeckx. *Temporal characteristics of ecohydrological variables in an intensively monitored wetland*. European Geosciences Union, General Assembly 2007, Vienna Austria, April 15-20, 2007. (Poster presentation in session HS32: Climate-soil and vegetation interactions in ecological-hydrological processes)

J. Peters, N. E. C. Verhoest, R. Samson, and B. De Baets. *Uncertainty propagation in vegetation distribution models based on ensemble learning*. Hydropredict2008: International Interdisciplinary Conference on Predictions for Hydrology, Ecology and Water Resources Management Using Data and Models to Benefit Society, Prague, Czech Republic, September 15-18, 2008. (Oral keynote presentation)

Deelname (passief) aan conferenties en workshops

BioScope-IT Workshop, Clustering & Classification in the Life Sciences. January 17, 2008, SAS Institute Tervuren, Belgium.

Deelname aan internationale wetenschappelijke projecten

July 2006 Deelname aan een ecohydrologisch project in Zuid-Afrika (Fund for Scientific Research, Flanders. Grant G.0443.05 by the Flemish Interuniversity Council.)

March 2005 Veldprospectie in het kader van een bilateraal project (Bilateral Scientific and Technological Cooperation between Flanders and Chile (Project BIL 01/04)) 'Ecohydrological monitoring and modelling in managed and unmanaged forest ecosystems of southern Chile'.

January 2004 Workshop en veldcampagne in het kader van een bilateraal project (Bilateral Scientific and Technological Cooperation between Flanders and Chile (Project BIL 01/04)) ‘Ecohydrological monitoring and modelling in managed and unmanaged forest ecosystems of southern Chile’.

September 23, 2008