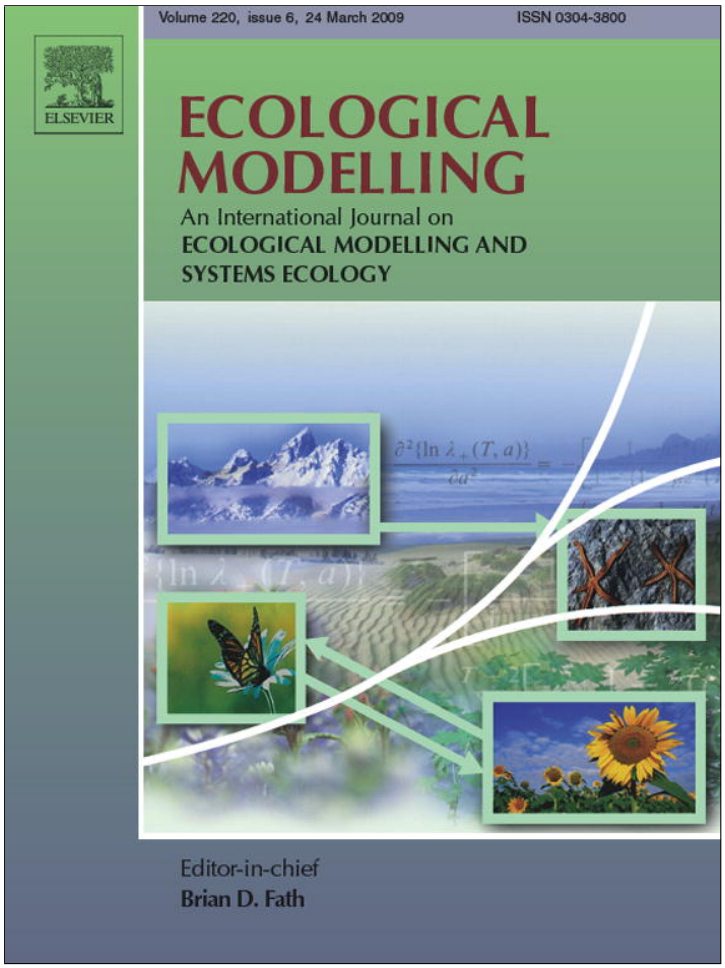

Ecological Modelling 220 (2009) 791–804

Contents lists available at ScienceDirect

Ecological Modelling

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

Uncertainty propagation in vegetation distribution models based on ensemble classifiers

Jan Peters[a,*], Niko E.C. Verhoest[a], Roeland Samson[b], Marc Van Meirvenne[c], Liesbet Cockx[c], Bernard De Baets[d]

[a] Department of Forest and Water Management, Ghent University, Coupure Links 653, B-9000 Gent, Belgium
[b] Department of Bioscience Engineering, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerpen, Belgium
[c] Department of Soil Management and Soil Care, Ghent University, Coupure Links 653, B-9000 Gent, Belgium
[d] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 10 April 2008
Received in revised form 8 December 2008
Accepted 12 December 2008
Available online 29 January 2009

Keywords:
Classification
Clustering
Distribution model
Ensemble learning
Random Forests
Spatial interpolation
Uncertainty
Vegetation type
Wetland

ABSTRACT

Ensemble learning techniques are increasingly applied for species and vegetation distribution modelling, often resulting in more accurate predictions. At the same time, uncertainty assessment of distribution models is gaining attention. In this study, Random Forests, an ensemble learning technique, is selected for vegetation distribution modelling based on environmental variables. The impact of two important sources of uncertainty, that is the uncertainty on spatial interpolation of environmental variables and the uncertainty on species clustering into vegetation types, is quantified based on sequential Gaussian simulation and pseudo-randomization tests, respectively. An empirical assessment of the uncertainty propagation to the distribution modelling results indicated a gradual decrease in performance with increasing input uncertainty. The test set error ranged from 30.83% to 52.63% and from 30.83% to 83.62%, when the uncertainty ranges on spatial interpolation and on vegetation clustering, respectively, were fully covered. Shannon's entropy, which is proposed as a measure for uncertainty of ensemble predictions, revealed a similar increasing trend in prediction uncertainty. The implications of these results in an empirical distribution modelling framework are further discussed with respect to monitoring setup, spatial interpolation and species clustering.

1. Introduction

The modelling of species and vegetation distributions based on their relationship with environmental variables 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 (Ferrier, 2002; Barry and Elith, 2006). The relationship between environmental variables and species or vegetation distributions has been described by a multitude of modelling techniques (Guisan and Zimmerman, 2000), among which generalized linear (GLM, McCullagh and Nelder, 1999), generalized additive (GAM, Hastie and Tibshirani, 1990; Yee and Mitchell, 1991) and machine learning techniques such as Random Forests (Benito Garzón et al., 2006; Lawler et al., 2006; Prasad et al., 2006; Araújo and New, 2007; Peters et al., 2007), neural networks (Foody, 1999; Özemsi et al., 2006; Westra and De Wulf, 2007), and support vector machines (Guo et al., 2005). are frequently applied. In attempting to describe complex distributional patterns, however, distribution modelling results will inevitably contain some degree of uncertainty (Barry and Elith, 2006), and the assessment of this uncertainty is gaining more and more attention in ecological modelling studies (e.g. Phillips and Marks, 1996; van Horssen et al., 2002; Larssen et al., 2007; Van Niel and Austin, 2007).

Uncertainty in vegetation distribution models is due to input data limitations, caused by spatial and temporal underrepresentation of observations, measurement and systematic errors on observations, missing of key environmental variables constraining the vegetation distribution, and subjective judgments such as judgment on the type of environmental variables vegetation is sensible to and their relative importance to classify vegetation types (Barry and Elith, 2006; Ray and Burgman, 2006). Furthermore, distribution modelling techniques introduce uncertainty by their inability to capture the entire complexity of ecological processes in relation to vegetation distributions. That is, 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.

Of the sources of error and uncertainty, this study exclusively investigates two important sources of uncertainty: (i) the uncertainty associated with the spatial interpolation of environmental

* Corresponding author. Tel.: +32 9 264 61 40; fax: +32 9 264 62 36.
E-mail address: jan.peters@ugent.be (J. Peters).

0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.12.022

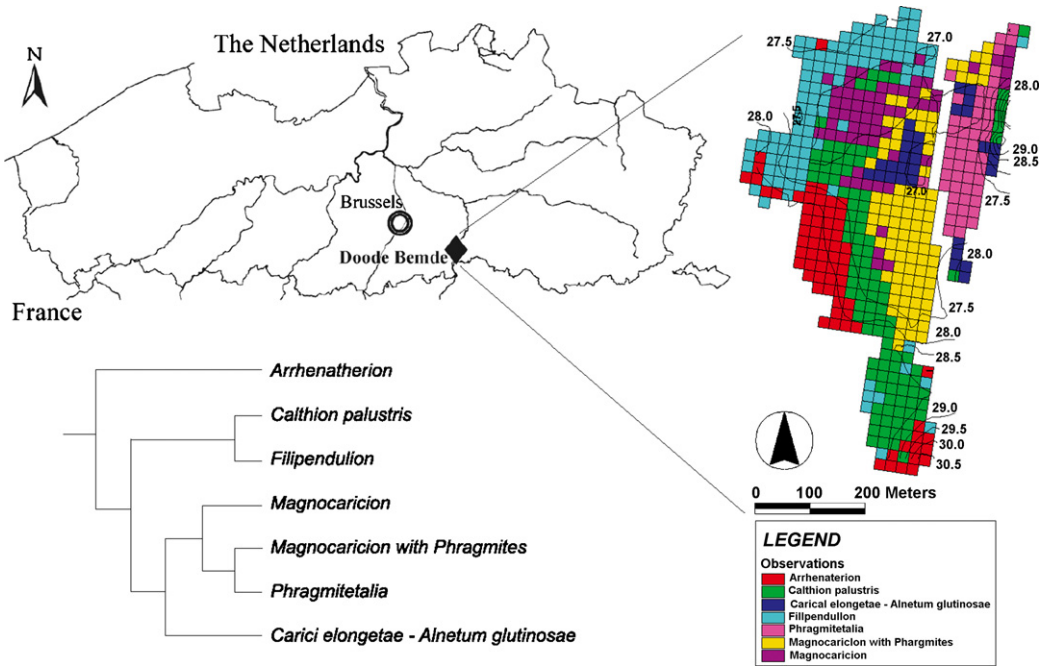

Fig. 1. Location of the study area. Detailed reproduction of site topography [m above reference level] and vegetation distributions (based on cluster dendrogram).

variables and (ii) the uncertainty associated with species clustering into vegetation types, and investigates its propagation in a vegetation distribution model based on Random Forests.

2. Materials and methods

2.1. Site description and monitoring network

A lowland river valley in Belgium called 'Doode Bemde' was the research area of this study (Fig. 1). Doode Bemde is a nature reserve which is part of a long term ecohydrological programme of the Research Institute of Nature and Forest (Belgium) (e.g. Huybrechts et al., 2000, 2002), and data from this long term investigation are applied in this study. The site is an alluvial floodplain mire in the middle course of the river Dijle, situated approximately 30 m above mean sea level. Mineral soils with silty texture and organic peat soils dominate the area. Large areas of the Doode Bemde are fed by nutrient rich groundwater at rates of approximately 3 mm day^{-1}(De Becker et al., 1999; De Becker and Huybrechts, 2000). The area is bordered by the river Dijle in the west, the Molenbeek, a tributary of the Dijle, in the north and the valley slope with a number of permanent springs in the east (De Becker et al., 1999). The climatic conditions at the site are typically temperate, with an average annual precipitation of approximately 800 mm distributed evenly over the year (Verhoest et al., 1997; De Jongh et al., 2006), an average annual pan evaporation of 450 mm, and an average annual air temperature of 9.8 °C (Van Herpe and Troch, 2000).

During the summer of 1993 plant species occurrences (presence/absence) were mapped in the study area. The total area of 21.08 ha was subdivided into 519 regular and adjacent grid cells of 20 m × 20 m, and referenced according to a local coordinate system. Presence/absence mapping was restricted to a selection of 56 plant species of which 45 were typically groundwater dependent (phreatophytes, *sensu* Londo, 1988) and 11 were indicative for the different vegetation types at the Doode Bemde.

A groundwater monitoring network consisting of 24 piezometers was installed in 1989, 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, and used to calculate the average groundwater depth (AGD) and amplitude of the groundwater depth (Ampli). 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}). Soil samples for organic matter content determination were taken at 59 locations at a depth of 0.05 m and 0.15 m and analysed using thermal destruction of the soil sample at 600 °C in a muffle furnace and expressed as a percentage (%). Management regime was assessed for each grid cell separately. Management regime was the only categorical variable in this study (the others are continuous), with four different regimes that could be distinguished:

- Yearly mowing in early summer, followed by grazing or mowing of the aftermath.
- Both yearly and cyclic mowing within the same grid cell.
- Cyclic mowing (once every 5–10 years) or not mown at all since at least 5, and up to 10 years.
- No management for at least 10 years.

The selection of the environmental variables is based on Peters et al. (2008a), in which they were identified as the ecologically most important variables in constraining the distribution of the wetland vegetation at the study site.

The spatio-temporal density of field observations varied between the different ecosystem compartments (Table 1). Management regime and soil organic matter content were described for every grid cell of the study area ($N = 519$) on a single occasion, while groundwater quality was measured twice and groundwater depth observations were made 26 times each year (every 2 weeks) in 24 piezometers ($n = 24$) scattered over the area. Brief summary statistics (mean, range, variance) of the environmental variables (Table 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. Further, groundwater quality as well as soil organic matter showed a high variability, and the study area could be concluded to comprise high variability in environmental conditions.

Table 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	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	mg L^{-1}	20.7	[5.3 76.1]	290.1
Vegetation	519	1	Management regime	–	–	–	–	–

2.2. Variation partitioning in species data

Spatial autocorrelation is a very general property of ecological variables (Legendre, 1993). Spatial structures observed in ecological communities arise from two independent processes (Legendre, 1993; Dray et al., 2006): (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 assemblage structure at surrounding localities, because of biotic processes such as growth, reproduction, mortality and migration. Variation partitioning (Borcard et al., 1992; Borcard and Legendre, 2002; Borcard et al., 2004) can be used to assess the importance of these two sources of spatial structure. Variation partitioning allows to code the spatial information into spatial variables which can be used in a direct gradient analysis such as partial canonical ordination (e.g. redundancy analysis, RDA Rao, 1964; van den Wollenberg, 1977 or canonical correspondence analysis, CCA Ter Braak, 1986), allowing for the partitioning of the total variation in the species data into the following four parts (Borcard et al., 1992):

- 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.
- The spatial structuring in the species data that is shared by the environmental data.
- The spatial patterns in the data that are not shared by the environmental data included in the analysis.
- The fraction of species variation explained neither by spatial nor by environmental variables.

2.3. Spatial interpolation using sequential Gaussian simulation

Point observations of environmental variables were spatially modelled using sequential Gaussian simulation (sGs, Goovaerts, 1997), mainly because of its ability to model local uncertainty. Additionally, sGs preserves the characteristic roughness in the data, not producing a smoothed estimate but a reproduction of the real variability (Alfaro, 1979). 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 cumulative distribution function (ccdf) of that variable $F_{\mathbf{u}_j}(z|(l)) = \text{Prob}\{Z(\mathbf{u}_j) \leq z|(l)\}$. To ensure reproduction of the z-semivariogram model, each ccdf is made conditional to local information $(|(l))$ not only including the observations but also the values simulated at previously visited locations. The sGs algorithm is nicely described by Bourennane et al. (2007), and Fagroud and Van Meirvenne (2002) provided a flow-chart. The sGs algorithm is available in the public domain (Deutsch and Journel, 1998).

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 $F^{-1}_{\mathbf{u}_j}(0.5)$ 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. 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, Goovaerts (2001) 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:

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

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

$$\xi_j(p) = \begin{cases} 1 & \text{if } F^{-1}_{\mathbf{u}_j}\left(\frac{1-p}{2}\right) < z(\mathbf{u}_j) \leq F^{-1}_{\mathbf{u}_j}\left(\frac{1+p}{2}\right), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

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 (Goovaerts, 2001) in which, for a series of probabilities p, the average width of the PIs that include the observed values are plotted. The average width $\bar{W}(p)$ is computed as:

$$\bar{W}(p) = \frac{1}{N_V \bar{\xi}(p)} \sum_{j=i}^{N_V} \xi_j(p) \cdot \left(F^{-1}_{\mathbf{u}_j}\left(\frac{1+p}{2}\right) - F^{-1}_{\mathbf{u}_j}\left(\frac{1-p}{2}\right) \right) \tag{3}$$

and should be as small as possible.

2.4. Species clustering

Cluster analysis of ecological data is an explicit way of identifying groups in data to find structures (Jongman et al., 1995). There are several clustering methods, and a major distinction can be made between divisive and agglomerative methods. In order to cluster species cover data into vegetation types, the TWINSPAN (Hill, 1979) program is frequently used in community ecology (Jongman et al., 1995). 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 (Gauch and Whittaker, 1981). So, 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 (Jaccard, 1901, 1912). 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).

2.5. Random forests

Random Forests (Breiman, 2001)(with capitals: referring to the technique) is an ensemble learning technique which generates many (k) classification trees (Breiman et al., 1984) that are aggregated, based on majority voting, to classify. 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 (Hansen and Salamon, 1990). Random Forests increases diversity among the classification trees by resampling the data with replacement, and by randomly changing the predictive variable sets over the different tree induction processes (technical details of the Random Forests algorithm are given on http://www.stat.berkeley.edu/~breiman/RandomForests/or in Cutler et al. (2007) and Peters et al. (2007, 2008b)). 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 predictive variable among a subset of m randomly selected predictive variables (Liaw and Wiener, 2002). 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 (without capitals: referring to the distribution model based on Random Forests). 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. Random Forests produces a limiting value of the generalization error (Breiman, 2001). As the number of trees increases, the generalization error always converges. The number of trees (k) needs to be set sufficiently high to ensure for this convergence. Consequently, Random Forests does not overfit. An upper bound for the generalization error can be obtained in terms of two parameters that measure how accurate the individual classification trees are and how diverse different classification trees are (Breiman, 2001): (i) the *strength* of each individual tree and (ii) the *correlation* between any two trees, where both parameters are not user-defined. 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 random forest error, whereas decreasing the correlation decreases the random forest error. Therefore m has to be optimized in order to get a minimal error.

Once appropriate parameter values are determined, Random Forests constructs an ensemble of k classification trees during 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:

$$P(c_j) = \frac{N_{c_j}}{N_{\text{tot}}} \tag{4}$$

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 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 output uncertainty. Therefore, the Shannon entropy measure (H, Shannon, 1948), which has been applied in other ecological modelling studies (e.g. Van Broekhoven et al., 2006; Ricotta and Anand, 2006), can be used:

$$H = -\frac{1}{\log_2 n}\sum_{j=1}^{n} P(c_j)\log_2 P(c_j) \tag{5}$$

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, i.e. the 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, i.e. the 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 provides a strong evidence for a certain vegetation type. Conversely, a value close to 1 indicates that, based on the environmental conditions, the random forest is not able to distinguish between the different vegetation types.

2.6. Evaluation of distribution modelling results

The lack of an independent test data set forced us to apply cross-validation for the random forest testing. In threefold cross-validation, a data set of N elements is randomly and uniformly split into three parts of $N/3$ elements, and three distribution models are constructed, each on data sets made up by two parts, and tested on the third part (Appendix A, Algorithm 1). Consequently, each element of data set is once used as a training instance, once as a test instance.

Several measures of classification accuracy are used throughout this study: the out-of-bag (oob) error, 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. The oob error is obtained during the Random Forests application and technical details and pseudo-code are given in Breiman (2001), Cutler et al. (2007) and Peters et al. (2007). Additionally, Cohen's κ test (Cohen, 1960), which corrects

the percent correctly classified for agreement that could be due to chance, was also used. 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. Finally, a threshold-independent evaluation using multi-class receiver operating characteristic (ROC) graphs was performed (Hosmer and Lemeshow, 2000; Fawcett, 2006). 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 (Provost and Domingos, 2001):

$$\mathrm{AUC}_{\mathrm{total}} = \sum_{c_j \in C} \mathrm{AUC}(c_j) \cdot w(c_j), \quad (6)$$

where $\mathrm{AUC}(c_j)$ is the area under the ROC curve for class c_j, and $w(c_j)$ a weighing factor based on class prevalences.

3. From field observations to a spatially distributed data set

3.1. Variation partitioning in species cover data

To quantify the spatial component of ecological variation at the Doode Bemde, variation partitioning (Borcard et al., 1992; Borcard and Legendre, 2002; Borcard et al., 2004) was applied to 21 grid cells within the study area. Within these grid cells 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) were constructed. 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 were assumed to show unimodal responses to the gradients in the study area, 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%.

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, such as 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 occurrences in relation with environmental 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 (Miller et al., 2007), which was beyond the study objectives.

Table 2
Summary of semivariogram models.

Variable	n	Model [a]	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
$SO\ SO_4^{2-}$	24	exp	0.14	1.11	319
SOM	59	sph	0.17	1.08	297

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 $|h| > 0$.
[a] Models ($\gamma(0) = 0$).

3.2. 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. Model parameters of the different semivariogram models are given in Table 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. 2, example of groundwater depth). Median values ($\hat{F}^{-1}(0.5)$) and conditional variances of these ccdfs were calculated. The conditional variance equaled 0 for grid cells where observations were made ($\hat{F}^{-1}(\cdot) = \mathrm{observedvalue}$). 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. 2). 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 us to apply leave-one-out cross-validation to assess local uncertainty (Van Meirvenne and Goovaerts, 2001). Data sets (containing all but one observations) 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 left-out elements. The median simulated values were plotted versus the observed values in scatter diagrams (Fig. 3) to investigate the local uncertainty. 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

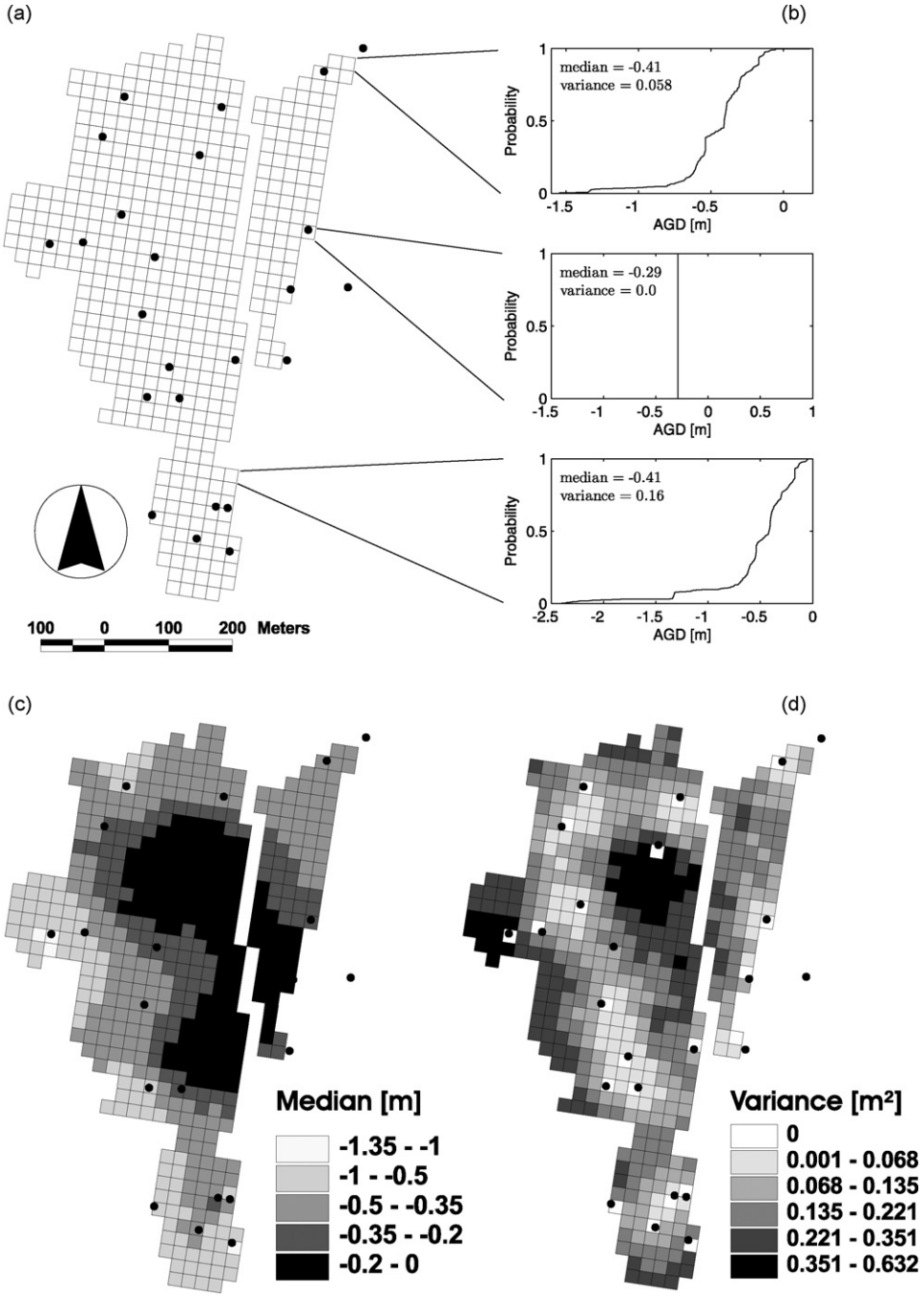

Fig. 2. Groundwater depths were monitored by piezometers (black dots, $n = 24$) scattered over the study area (a). Sequential Gaussian simulation using these observations resulted in 500 equiprobable groundwater depth realizations for each grid cell ($N = 519$). Empirical non-parametric conditional cumulative 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.

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 sGs 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 the text.

Furthermore, data sets with increasing uncertainty on the continuous environmental variables were constructed by applying Latin hypercube sampling (McKay et al., 1979; Imam and Conover, 1980) on the realizations of the sGs simulation. Latin hypercube sampling is a stratified random procedure that provides an efficient way of sampling variables from their cumulative probability distributions (Minasny and McBradney, 2006), and five different probability intervals were choosen to sample from. These probability intervals were symmetrical around probability 0.5, and are represented as $[0.5 - a; 0.5 + a]$ with

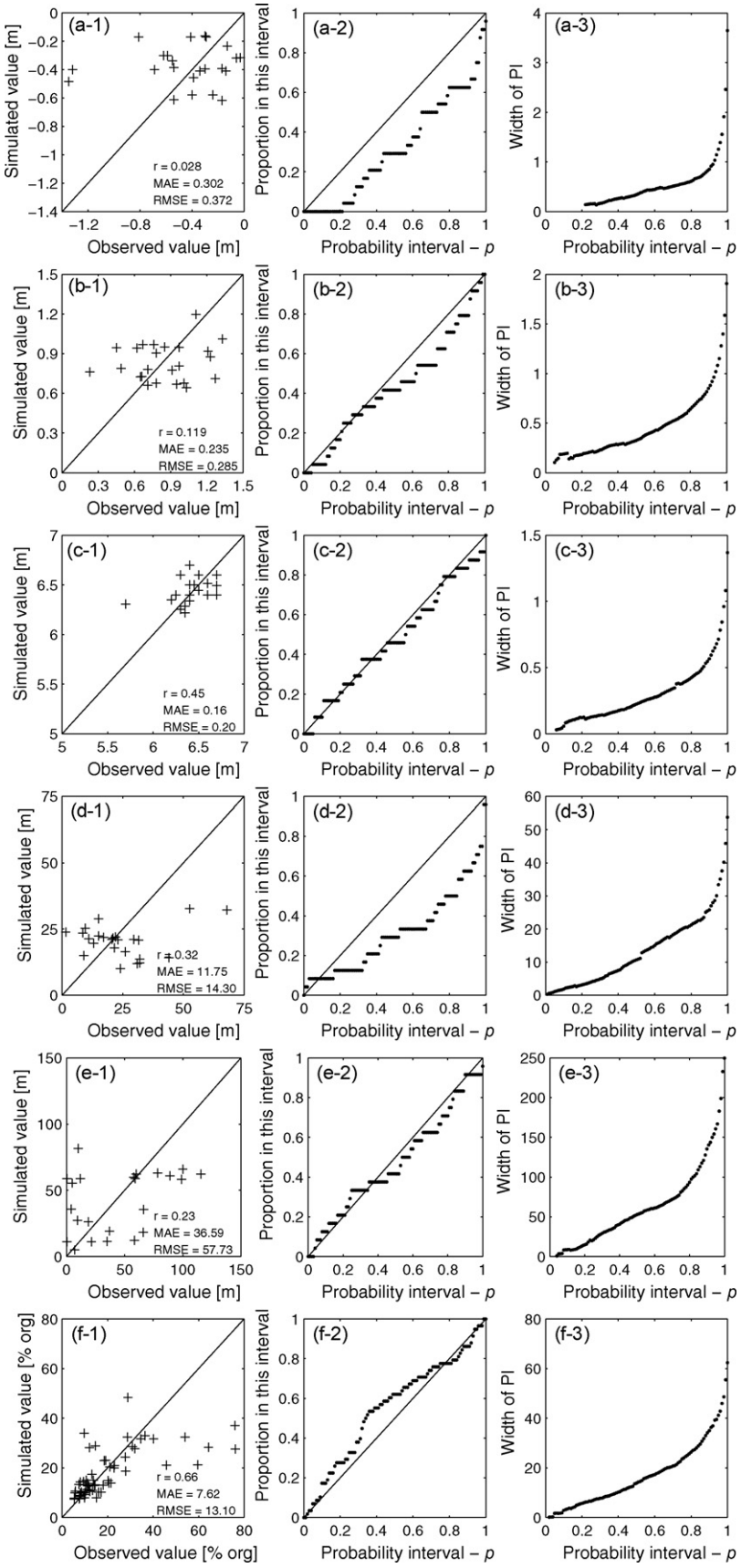

Fig. 3. Local uncertainty assessment by means of leave-one-out cross-validation: scatter diagram (1), accuracy plot (2) and precision plot (3) for the simulation results of the variables AGD (a), Ampli (b), pH (c), Cl^{-} (d), SO_4^{2-} (e), SOM (f).

Table 3
Jaccard index of similarity between the vegetation types in the Doode Bemde.

	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

Abbreviations: Ar, Arrhenatherion; Cp, Calthion palustris; Ce, Carici elongetae–Alnetum glutinosae; Fp, Filipendulion; Ph, Phragmitetalia; MP, Magnocaricion with Phragmites; Ma, Magnocaricion.

$a \in \{0.005, 0.05, 0.1, 0.25, 0.5\}$. From these probability intervals samples were drawn, using Latin hypercube sampling. The number of samples was made proportional to the width of the probability interval, and 1 sample was taken from $\hat{F}^{-1}([0.495; 0.505])$, 10 samples from $\hat{F}^{-1}([0.45; 0.55])$, 20 samples from $\hat{F}^{-1}([0.4; 0.6])$, 50 samples from $\hat{F}^{-1}([0.25; 0.75])$, and 100 samples from $\hat{F}^{-1}([0; 1])$. Each sample was linked with the categorical variables management and vegetation type, and as such data sets constructed were constructed, which are represented as $L_a^e(s)$, where e refers to the uncertainty on environmental variable estimations which is quantified by these data sets, and s to the number of Latin hypercube samples drawn from the probability interval $[0.5 - a; 0.5 + a]$.

3.3. Uncertainty on species clustering

Based on the species cover data, TWINSPAN (Hill, 1979) was applied in order to define vegetation types. Seven different vegetation types were distinguished at the study site. A simplified representation of the TWINSPAN dendrogram and the spatial distribution of the seven different vegetation types can be seen in Fig. 1. A more detailed description of these vegetation types is given by De Becker et al. (1999) and Peters et al. (2007).

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 have to be made and influence the solution (Ter Braak et al., 2003). Additionally, the stability of the TWINSPAN solution is often of concern (Vangroenewoud, 1992; Oksanen and Minchin, 1997; Ter Braak et al., 2003).

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 3 for the seven different vegetation types. The values of the diagonal elements in Table 3 are a measure of similarity between grid cells of the same vegetation type, and do not necessarily equal 1 since species composition between grid cells clustered in the same vegetation type may differ considerably. 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 have lower values. Between the different vegetation types, marked differences in similarity could be observed. *Magnocaricion with Phragmites* has high similarities with *P* hragmitetalia and *Magnocaricion*. Between the other vegetation types, similarities are generally lower, but nevertheless differences can be observed. *Arrhenatherion* for example, has twice as much species in common with *Filipendulion* than with *M* agnocaricion.

Based on this analysis, six 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 3). This strategy reflects the likelihood of erroneous clustering of a grid cell based on its species composition. An *Arrhenatherion* grid cell for example, had on average approximately twice as much species in common with *F* ilipendulion as 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_b^v where superscript v refers to the uncertainty in species clustering into vegetation types and subscript b to the percentage of pseudo-randomized elements used to quantify this source of uncertainty.

4. Model construction, calibration and evaluation

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. Both parameters have to be calibrated to minimize the random forest error. In addition to the built-in out-of-bag model generalization error estimator, threefold cross-validation was applied for the random forest testing. Consequently, each measurement vector $\mathbf{x}_i$ was classified by k trees as a unique vegetation type, and these results were used to compute the final classification based on majority voting. Oob error and test set error were averaged over the three random forests in threefold cross-validation using different values of m. Fig. 4 shows convergence of the random forests constructed with different numbers of m ($m = 1$ (minimal value), $m = 3$ (optimal value, in accordance with Breiman's rule of thumb ($\sqrt{\text{number of variables}}$, Breiman, 2001), and $m = 7$ (maximal value)) when more trees are added (i.e. k increases). Based on this calibration, the values 1000 and 3 were used for the two user-defined parameters k and m, respectively.

Using these parameter values, the random forest performed 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 (Cohen, 1960) of 0.633 was calculated, indicating a substantial agreement between observations and predictions. A threshold-independent evaluation using multi-class ROC graphs was performed. For each vegetation type a ROC curve was produced (Fig. 5) and its AUC was calculated

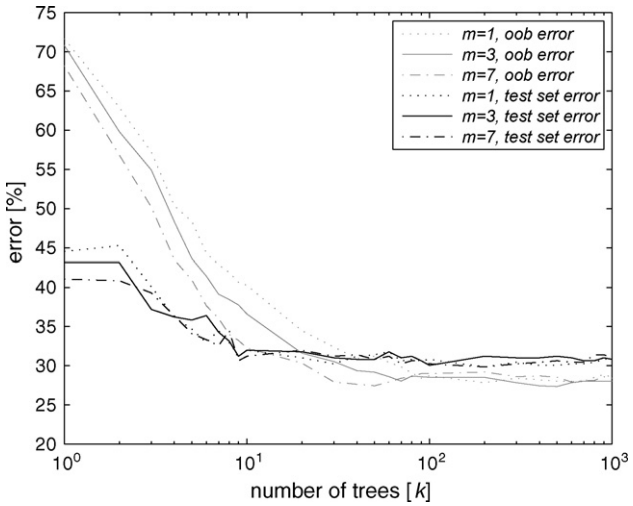

Fig. 4. Out-of-bag (oob) error and test set error converge when more trees are added to the random forest (when k increases). The numbers of variables (m) used to split the nodes are $m = 1$, 3 and 7. Average error values of the threefold cross-validated random forest are plotted.

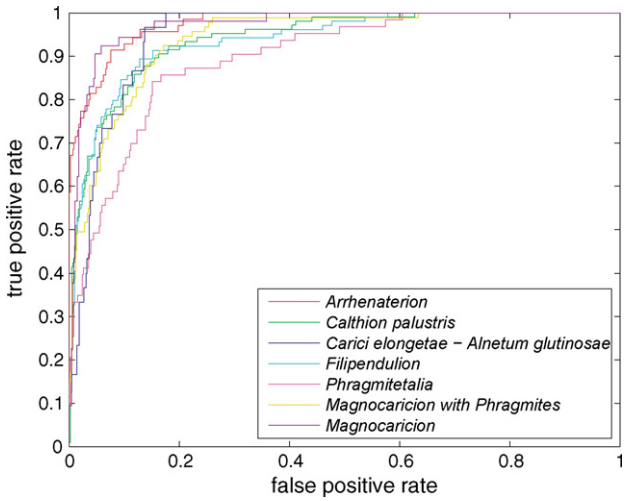

Fig. 5. Receiver operating characteristic (ROC) curves visualizing the classification performances of the threefold cross-validated random forest for the seven vegetation types.

and averaged over the different classes using class weights based on class prevalences in the dataset (Eq. (6)). The AUC_{total} value equaled 0.943 and the random forest was concluded to perform well.

The random forest output for each grid cell is a discrete probability distribution over the seven vegetation classes (see Eq. (4)). Looking into this probability distribution by means of Shannon's entropy measure H (Eq. (5)) allowed to gain some information on the output uncertainty. H values are between the minimal value 0 and the maximal value 1. 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. 6(a)), a decrease in frequency counts could be seen with increasing H values. This means that the random forest output distribution was generally quite narrow, with a clear dominance of one, two or — to a lower extent — three different vegetation types.

5. Uncertainty propagation to the modelling results

Model input data sets inevitably contain uncertainties. Two major sources of uncertainty, namely: (i) uncertainty associated with spatial interpolation of environmental variables and (ii) uncertainty on species clustering, and their propagation to the modelling results are assessed in this study.

5.1. Uncertainty on spatial interpolation of environmental variables

Threefold cross-validation was applied to construct random forests and to test them on the propagation of uncertainty due to uncertain environmental variables. The construction was based on two folds from the reference data set L. Grid cells of the third fold were identified by means of their local coordinates, and drawn from the Latin hypercube test data sets ($L_a^e(s)$). Random forest testing was repeated for each sample (s) from each probability interval $[0.5 - a; 0.5 + a]$, with $a \in \{0.005, 0.05, 0.1, 0.25, 0.5\}$ (Appendix A, Algorithm 2). So, each element of $L_a^e(s)$ was once used as test element. For each uncertainty level a, s ($s = 2 \cdot a \times 100$) probabilities of occurrence for all seven vegetation types were modelled for each grid cell in the study area. Results indicated increasing test set errors, and decreasing κ and AUC_{total} values when a increased (Table 4). However, the increase in test set error and decrease in κ and AUC_{total} was limited when $a \leq 0.1$, and much more pronounced when this threshold was exceeded. Indeed, the vegetation type of

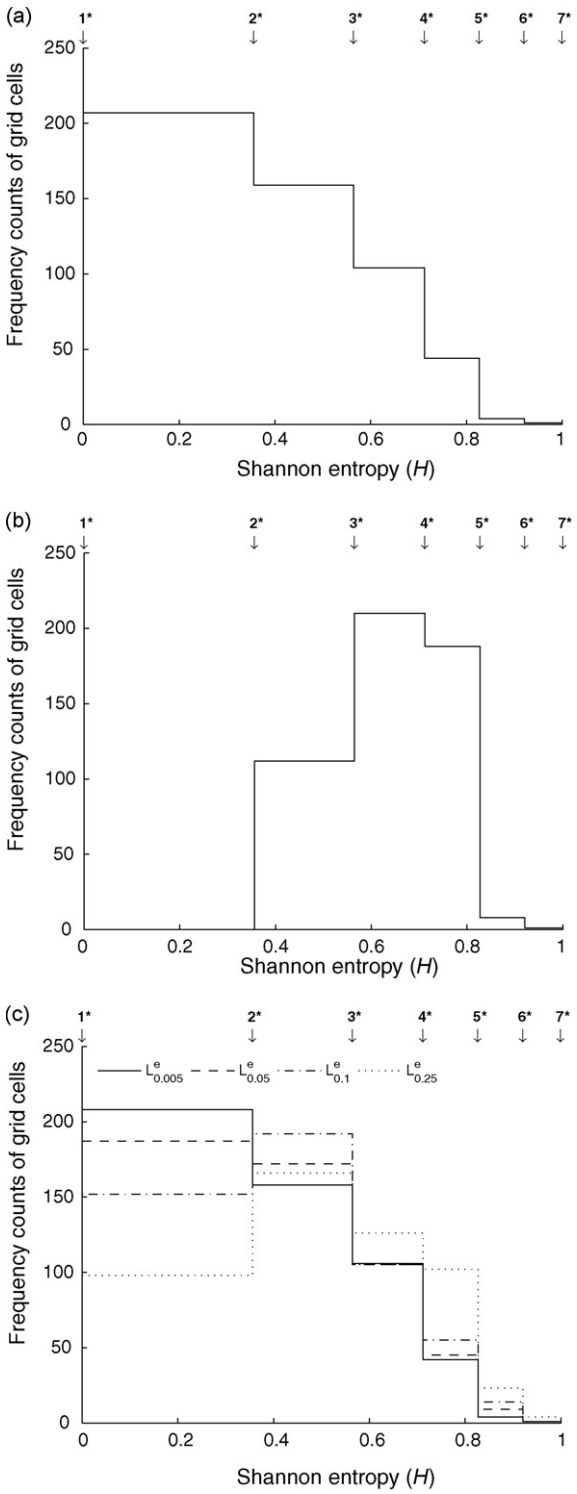

Fig. 6. Histogram of frequency counts of the Shannon entropy (H) values of the entire study site ($N = 519$) for the random forest cross-validated on L (a), and tested on the Latin hypercube samples (averaged) $L^e_{0.50}$ (b), and the gradually deviating test data sets L^e_a with $a = 0.005$, $a = 0.05$, $a = 0.1$ and $a = 0.25$ (c). Legend: **j*** indicates the values of H obtained when a grid cell is classified as **j** vegetation types with equal probability of occurrence ($P(c_j) = 1/\mathbf{j}$).

Table 4
Uncertainty on environmental variables propagating to the random forest results. Results are averaged over the number of samples for $\tilde{L}^e_{0.05}$, $\tilde{L}^e_{0.1}$, $\tilde{L}^e_{0.25}$ and $\tilde{L}^e_{0.5}$.

Data set	oob error (%)	Test set error (%)	Cohen's κ	AUC_{total}	Average H
L	28.03	30.83	0.633	0.943	0.420
$\tilde{L}^e_{0.005}$	28.03	30.64	0.635	0.943	0.422
$\tilde{L}^e_{0.05}$	28.03	31.25	0.628	0.939	0.438
$\tilde{L}^e_{0.1}$	28.03	31.95	0.619	0.932	0.470
$\tilde{L}^e_{0.25}$	28.03	39.76	0.523	0.901	0.550
$\tilde{L}^e_{0.5}$	28.03	52.63	0.367	0.828	0.661

only 1.12% of the test elements drawn from the [0.4; 0.6] probability interval (approximately 6 test grid cells) were incorrectly predicted by the random forest which was constructed and calibrated on reference (median, see Section 3.2) values. Contrarily, the evaluation statistics for the Latin hypercube samples covering the entire probability interval ($a = 0.5$) indicated inaccurate performance; from the 100 (s) test data sets containing 519 elements, on average only 229.7 elements (47.37%) were classified correctly (compared to 69.17% on the reference training set), and a κ value of 0.367 and AUC_{total} value of 0.828 were obtained.

A more detailed investigation of these modelling results was made by a grid-wise comparison of variances (Fig. 7). It was hypothesised that grid cells with low variances in sGs outcomes for the continuous environmental variables (i.e. grid cells where observations were made, for which simulated values equaled the observed value, $\hat{F}^{-1}(\cdot) =$ observedvalue, as an extreme example) have a low variance in the maximal probability of occurrence. Therefore six scatter plots were constructed, one for each continuous variable, in which the variance of the simulation results is plotted against the variance of $P(c_{max})$ for all 100 Latin hypercube test runs. Four different groups were created within each plot based on classification accuracy, and by applying Spearman's ρ, correlations between variances were calculated. Significant positive correlations at the

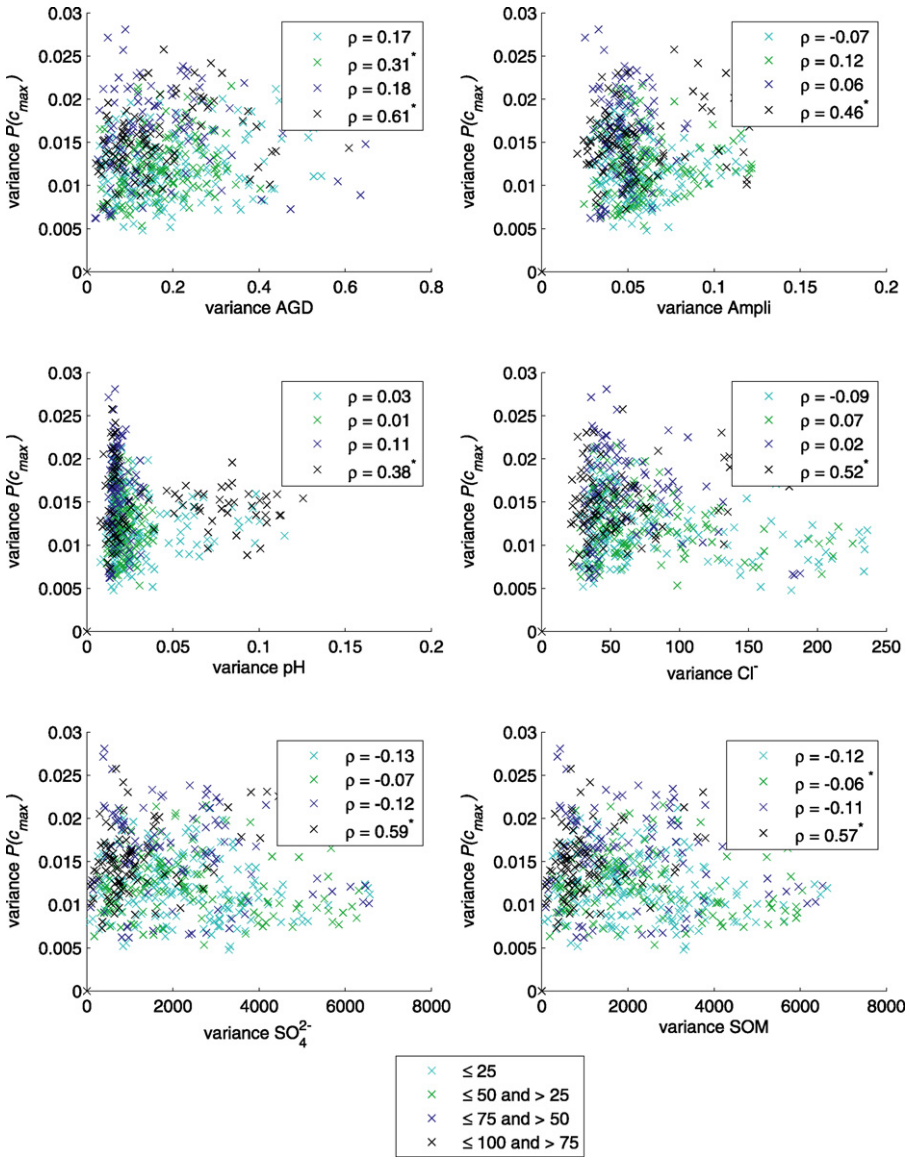

Fig. 7. Variable variance versus variance in modelled probability of occurrence of the predicted vegetation type ($P(c_{max})$) when the random forest is applied to 100 Latin hypercube test data sets drawn from the entire probability distributions. Different colours group the scatter points ($N = 519$)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

Table 5
Uncertainty on species clustering propagating to the random forest results.

Data set	oob error (%)	Test set error (%)	Cohen's κ	AUC_{total}	Average H
L	28.03	30.83	0.633	0.943	0.420
L_b^v					
$b = 1\%$	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

0.05 significance level were found for grid cells that were classified correctly in $>$ 75 of the 100 test 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.

The entropy (H) of the random forest output for all grid cells i, averaged over all s Latin hypercube samples $L_a^e(s)$ for each uncertainty level a, resulted in a histogram of frequency counts. The first subfigure (Fig. 6(a)) shows the entropy distribution among the 519 grid cells for the modelling based on the reference data set, while the second subfigure shows the average results when modelling was tested on $L_{0.5}^e(s)$ (Fig. 6(b)). Where a maximum was found at entropy values between 0 and 0.356 for the reference set, a maximum between 0.565 and 0.712 was observed when the probability of occurrence of the vegetation types was based on highly uncertain environmental information. Grid cells were mostly classified as three or four different vegetation types with similar probabilities of occurrence. None of the grid cells was classified with a H value <0.356. In comparison with the histogram based on the cross-validated results of the random forest on the reference data set, a clear shift toward higher H values was observed, indicating that uncertainties on the spatial interpolation are propagated to the random forest results. This trend was confirmed by the moderately uncertain test data sets (Fig. 6(c)), where a shift toward higher entropy values could be observed for increasing values of a, reflecting the increasing uncertainty of the random forest.

5.2. Uncertainty on species clustering

The data sets with pseudo-randomizations in the response variable (L_b^v) were used for random forest construction and testing (Appendix A, Algorithm 3). The reason why the calibrated random forest constructed on the reference data set L was not used here, is that Random Forests constructs its classifiers taking response variables into account (supervised learning), and hence the uncertainty related to species clustering should be taken into account during model construction as well.

Random forests constructed on data sets with an increasing proportion of elements pseudo-randomized in the response variable, showed increasing oob errors (Table 5): 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 performances did deteriorate gradually with increasing percentages of the elements pseudo-randomized. For the other evaluation statistics, similar conclusions hold.

6. Discussion and conclusion

Vegetation distribution models tend to describe vegetation patterns based on environmental variables. A variety of uncertainty sources can, however, affect vegetation distribution modelling results. A first source of uncertainty under investigation was the uncertainty associated with the spatial interpolation of environmental variables. To preliminary assess the relative importance of the environmental variables and their spatial variation in constraining the wetland vegetation at the study site, variation partitioning was applied. 20.8% of the variation in vegetation distribution could be explained by the environmental variables as such, while 37.6% could be attributed to environmental gradients (i.e. the spatial variability in environmental conditions). Most frequently in distribution modelling, however, area covering estimates of environmental gradients are obtained by geostatistical interpolation techniques, based on a (limited) number of observations. These estimates inevitably contain a certain degree of uncertainty.

Sequential Gaussian simulation was applied to estimate the environmental gradients based on point observations, thereby enabling the quantification of local uncertainty. Simulation results were not accurate for most of the environmental variables, and conditional cumulative density functions showed a high variability for most grid cells. The two main reasons for this inaccuracy are the spatial underrepresentation of observations and the presence of strong environmental gradients in the conditioning data. Groundwater observations were already made in a quite densely arranged piezometer network (approximately 1.1 piezometer/ha), and increasing the piezometer density may not be feasible from a practical point of view. A potential way to increase interpolation accuracy without increasing monitoring densities is the use of secondary data (e.g. topography) for spatial interpolation. One could argue only to include the observed point measurements in order reduce environmental uncertainty. However, learning techniques (as other statistical techniques) on which distribution modelling is based, demand for environmental data covering a substantial environmental amplitude of the vegetation types for the model to gain generalizabilty and applicability. The inclusion of an environmental gradient in the distribution data is therefore necessary.

The conditional cumulative density functions generated by sGs were further used to investigate the propagation of uncertain environmental descriptions to the distribution modelling results. The conditional cumulative density functions were therefore gradually sampled to construct data sets with increasing uncertainty. Random Forests was applied to these data sets, and evaluation measures indicated a decreasing performance when an uncertainty threshold was exceeded. In this study, the uncertainty threshold was identified as the [0.4; 0.6] probability interval; if variable values ranged between $\hat{F}^{-1}(0.4)$ and $\hat{F}^{-1}(0.6)$, the random forests performed satisfactorily. The fact that the model performs well with variable values within a certain range from training values, is probably the reason why the model constructed and tested on the reference data set performed well (see Section 4), given the inaccuracies on continuous variable estimates (see Section 3.2).

This uncertainty assessment emphasized that environmental variables with limited uncertainty are important 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.

A second source of uncertainty under investigation was associated with species clustering into vegetation types. Vegetation type delineation based on species composition is a commonly used practice in ecology, and therefore it is highly relevant to assess the effect of introduced uncertainty in a vegetation distribution modelling context. Random Forests was applied to pseudo-randomized

data sets accounting for the likelihood of erroneuos species clustering. This assessment allowed to get insight in the decreasing performance with increasing uncertainty on the response variable, and results stressed the importance of accurate species mapping and vegetation type determination. 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 (Guisan and Zimmerman, 2000). However, dominant species are not necessarily the most ecologically relevant in distribution modelling. Furthermore, as vegetation types are frequently used in nature conservation, management and legislation (e.g. Landolt, 1994; Wood, 2000; Dias et al., 2004; Kleinod et al., 2005), the application of vegetation distribution models will remain important.

Acknowledgements

The authors wish to thank the special research fund (BOF) (project number 011/015/04) of Ghent University (Belgium), and the Fund for Scientific Research–Flanders (operating and equipment grant 1.5.108.03). We are also grateful to the Research Programme on Nature Development (projects VLINA 96/03 and VLINA 00/16) of the Flemish Government. The reviewers are acknowledged for their valuable comments.

Appendix A

Algorithm 1. Pseudo-code for random forest construction and testing using threefold cross-validation.

partition the reference data set $L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}$ *with* $N = 519$ *into 3 disjoint test data sets* T_1, T_2 *and* T_3*;*

for $i = 1 : 3$ ***do***

use $S_i = L \setminus T_i$ *to construct random forest* RF_i*;*

calculate the out-of-bag-error;

save RF_i*;*

apply the random forest to test data set T_i*;*

calculate the test set error;

save $P(c_j) = N_{c_j}/N_{\text{tot}}$ *(Eq. (4)) for all elements of* T_i*;*

end

calculate test statistics κ*,* AUC_{total} *and* H

Algorithm 2. Pseudo-code for random forest testing with gradually deviating test data sets.

for $a \in \{0.005, 0.05, 0.1, 0.25, 0.5\}$ ***do***

  for $s = 1 : 2 \cdot a \times 100$ ***do***

    for $i = 1 : 3$ ***do***

    use the partitioning of Algorithm 1 to partition the data set $L_a^e(s)$ *into 3 disjoint test data sets* $T_{a,1}(s), T_{a,2}(s)$ *and* $T_{a,3}(s)$*;*

    apply the saved random forest (RF_i) *to test data set* $T_{a,i}(s)$*;*

    calculate the test set error;

    save $P_{a,s}(c_j) = N_{a,s,c_j}/N_{\text{tot}}$ *(Eq. (4)) for all elements of* $T_{a,i}(s)$*;*

end

calculate test statistics $\kappa_a(s)$, $AUC_{a,\mathrm{total}}(s)$ *and* $H_a(s)$;

end

end

average test statistics

Algorithm 3. Pseudo-code for random forest testing with uncertainty on species clustering.

for $b \in \{1\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$ ***do***

use the partitioning of Algorithm 1 to partition the data set L_b^v *into 3 disjoint test data sets* $T_{b,1}, T_{b,2}$ *and* $T_{b,3}$;

for $i = 1:3$ ***do***

use $S_i = L_b^v \setminus T_{b,i}$ *to construct random forest* $RF_{b,i}$;

calculate the out-of-bag-error;

apply the random forest ($RF_{b,i}$) *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. (4)) for all elements of* $T_{b,i}$;

end

calculate test statistics κ, AUC_{total} *and* H;

end

References

Alfaro, M., 1979. Étude de la robustesse des simulations de fonctions aléatoires. Doctoral Thesis. E.N.S. des Mines de Paris.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22 (1), 42–47.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43, 413–423.

Benito Garzón, M., Blazek, R., Neteler, M., Sánchez de Dios, R., Sainz Ollero, H., Furlanello, C., 2006. 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.

Borcard, D., Legendre, P., 2002. All-scale analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153 (1–2), 51–68.

Borcard, D., Legendre, P., Avois-Jacquet, C., Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85 (7), 1826–1832.

Borcard, D., Legendre, P., Drapeau, P., 1992. Partialling out the spatial component of ecological variation. Ecology 73 (3), 1045–1055.

Bourennane, H., King, D., Couturier, A., Nicoullaud, B., Mary, B., Richard, G., 2007. 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.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall, New York.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler Jr., J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

De Becker, P., Hermy, M., Butaye, J., 1999. Ecohydrological characterization of a groundwater-fed alluvial floodplane mire. Applied Vegetation Science 2, 215–228.

De Becker, P., Huybrechts, W., 2000. De Doode Bemde—Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium (in Dutch).

De Jongh, I.L.M., Verhoest, N.E.C., De Troch, F.P., 2006. Analysis of a 105-year time series of precipitation observed at Uccle, Belgium. International Journal of Climatology 26, 2023–2039.

Deutsch, C., Journel, A., 1998. GSLIB: Geostatistical Software Library and User's Guide. In: Applied Geostatistics Series, 2nd ed. Oxford University Press, Oxford, UK.

Dias, E., Elias, R.B., Nunes, V., 2004. Vegetation mapping and nature conservation: a case study in Terceira Island (Azores). Biodiversity and Conservation 13 (8), 1519–1539.

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196 (3–4), 483–493.

Fagroud, M., Van Meirvenne, M., 2002. Accounting for soil autocorrelation in the design of experimental trials. Soil Science Society of America Journal 66, 1134–1142.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.
Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology 51, 331–363.
Foody, G.M., 1999. Applications of self organizing feature map neural network in community data analysis. Ecological Modelling 120 (2–3), 97–107.
Gauch, H.G., Whittaker, R.H., 1981. Hierarchical classification of community data. Journal of Ecology 69, 537–557.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.
Goovaerts, P., 2001. Geostatistical modeling of uncertainty in soil science. Geoderma 103, 3–26.
Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135 (2–3), 147–186.
Guo, Q.H., Kelly, M., Graham, C.H., 2005. Support vector machinespredicting distribution of sudden Oak death in Calefornia. Ecological Modelling 182 (1), 75–90.
Hansen, L., Salamon, P., 1990. Neural network ensembles. IEEE Transaction on Pattern Analysis and Machine Intelligence 12, 993–1001.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
Hill, M.O., 1979. TWINSPAN—A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes. Cornell University, Ithaca.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. Wiley, New York.
Huybrechts, W., Batelaan, O., De Becker, P., Joris, I., Van Rossum, P., 2000. Ecohydrologisch onderzoek waterrijke vallei-ecosystemen (VLINA 96/03). Instituut voor Natuurbehoud, Brussels (in Dutch).
Huybrechts, W., De Bie, E., De Becker, P., Wassen, M., Bio, A., 2002. Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, ITORS-VL (VLINA 00/16). Instituut voor Natuurbehoud, Brussels (in Dutch).
Imam, R.L., Conover, W.J., 1980. Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. Communications in Statistics Theory and Methods A9, 1749–1874.
Jaccard, P., 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin del la Société Vaudoise des Sciences Naturelles 37, 241–272.
Jaccard, P., 1912. The distribution of the flora of the alpine zone. New Phytologist 11, 37–50.
Jongman, R.H.G., ter Braak, C.J.F., Tongeren, O.F.R.V. (Eds.), 1995. Data Analysis in Community and Landscape Ecology, 2nd ed. Cambridge University Press, Cambridge.
Kleinod, K., Wissen, M., Bock, M., 2005. Detecting vegetation changes in a wetland area in Northern Germany using earth observation and geodata. Journal of Nature Conservation 13, 115–125.
Landolt, E., 1994. Vegetation mapping and nature conservation in Switzerland. Plant Ecology 110 (1), 19–23.
Larssen, T., Høgåsen, T., Cosby, B.J., 2007. Impact of time series data on calibration and prediction uncertainty for a deterministic hydrogeochemical model. Ecological Modelling 207 (1), 22–33.
Lawler, J.J., White, D., Neilson, R.P., Blaustein, A.R., 2006. Predicting climate-induced range shifts: model differences and model reliability. Global Change Biology 12, 1568–1584.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2/3, 18–22.
Londo, G., 1988. Nederlandse Freatophyten. Pudoc, Wageningen (in Dutch).
McCullagh, P., Nelder, J.A., 1999. Generalized Linear Models, 2nd ed. Chapman and Hall, Boca Raton.
McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables n the analysis of output from a computer code. Technometrics 21, 239–245.
Miller, J., Franklin, J., Aspinall, R., 2007. Incorporating spatial dependence in predictive vegetation models. Ecological Modelling 202 (3–4), 225–242.
Minasny, B., McBradney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32, 1378–1388.
Oksanen, J., Minchin, P.R., 1997. Instability of ordination results under changes in input data order: explanations and remedies. Journal of Vegetation Science 8 (3), 447–454.
Özemsi, S.L., Tan, C.O., Özemsi, U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 195 (1–2), 83–93.
Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207 (2–4), 304–318.
Peters, J., Verhoest, N.E.C., Samson, R., Boeckx, P., De Baets, B., 2008a. Wetland vegetation distribution modelling for the identification of constraining environmental variables. Landscape Ecology 23 (9), 1049–1065.
Peters, J., De Baets, B., Samson, R., Verhoest, N.E.C., 2008b. Modelling groundwater-dependent vegetation patterns using ensemble learning. Hydrology and Earth System Sciences 12, 603–613.
Phillips, D.L., Marks, D.G., 1996. Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models. Ecological Modelling 91 (1–3), 213–229.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
Provost, F., Domingos, P., 2001. Well-trained PETs: Improving probability estimation trees. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, NY.
Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhya A 26, 329–358.
Ray, N., Burgman, M.A., 2006. Subjective uncertainties in habitat suitability maps. Ecological Modelling 195 (3–4), 172–186.
Ricotta, C., Anand, N., 2006. Spatial complexity of ecological communities: bridging the gap between probabilistic and non-probabilistic uncertainty measures. Ecological Modelling 197 (1–2), 59–66.
Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656.
Ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67 (5), 1167–1179.
Ter Braak, C.J.F., Hoijtink, H., Akkermans, W., Verdonschot, P.F.M., 2003. Bayesian model-based cluster analysis for predicting macrofaunal communities. Ecological Modelling 160 (3), 235–248.
Vangroenewoud, H., 1992. The robustness of correspondence, detrended correspondence, and TWINSPAN analysis. Journal of Vegetation Science 3 (2), 239–246.
Van Broekhoven, E., Adriaenssens, V., De Baets, B., Verdonschot, P.F.M., 2006. Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. Ecological Modelling 198 (1–2), 71–84.
van den Wollenberg, A.L., 1977. Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42, 207–219.
Van Herpe, Y., Troch, P.A., 2000. Spatial and temporal variations in surface water nitrate concentrations in a mixed land use catchment under humid temperate climati conditions. Hydrological Processes 14, 2439–2455.
van Horssen, P.W., Pebesma, E.J., Schot, P.P., 2002. Uncertainties in spatially aggregated predictions from a logistic regression model. Ecological Modelling 154 (1–2), 93–101.
Van Meirvenne, M., Goovaerts, P., 2001. Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. Geoderma 102, 75–100.
Van Niel, K.P., Austin, M.P., 2007. Predictive vegetation modelling for conservation: impact of error propagation from digital elevation data. Ecological Applications 17 (1), 266–280.
Verhoest, N.E.C., Troch, P.A., De Troch, F.A., 1997. On the applicability of Barlett–Lewis rectangular pulses models in the modeling of design storms at a point. Journal of Hydrology 202, 108–120.
Westra, T., De Wulf, R.R., 2007. Monitoring Sahelian floodplains using Fourier analysis of MODIS time-series data and artificial neural networks. International Journal of Remote Sensing 28 (7–8), 1595–1610.
Wood, B., 2000. 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.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2, 587–602.