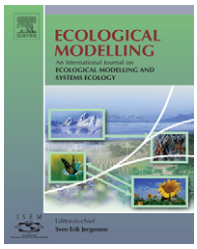
available at www.sciencedirect.com

ScienceDirect

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



Random forests as a tool for ecohydrological distribution modelling

Jan Peters[a,*], *Bernard De Baets*[b], *Niko E.C. Verhoest*[a], *Roeland Samson*[c], *Sven Degroeve*[b], *Piet De Becker*[d], *Willy Huybrechts*[d]

[a] *Department of Forest and Water Management, Ghent University, Coupure links 653, B-9000 Gent, Belgium*
[b] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium*
[c] *Department of Applied Ecology and Environmental Biology, Ghent University, Coupure links 653, B-9000 Gent, Belgium*
[d] *Research Group Ecohydrology and Water Systems, Institute of Nature Conservation, Kliniekstraat 25, B-1070 Brussels, Belgium*

ARTICLE INFO

Article history:
Received 17 March 2006
Received in revised form
30 April 2007
Accepted 23 May 2007
Published on line 20 July 2007

Keywords:
Vegetation model
Random forest
Classification tree
Logistic regression
Generalized linear model
Ecohydrology

ABSTRACT

An important issue in ecohydrological research is distribution modelling, aiming at the prediction of species or vegetation type occurrence on the basis of empirical relations with hydrological or hydrogeochemical habitat conditions. In this study, two statistical techniques are evaluated: (i) the widely used multiple logistic regression technique in the generalized linear modelling framework, and (ii) a recently developed machine learning technique called 'random forests'. The latter is an ensemble learning technique that generates many classification trees and aggregates the individual results. The two different techniques are used to develop distribution models to predict the vegetation type occurrence of 11 groundwater-dependent vegetation types in Belgian lowland valley ecosystems based on spatially distributed measurements of environmental conditions. The spatially distributed data set under investigation consists of 1705 grid cells covering an area of 47.32 ha. After model construction and calibration, both models are applied to independent test data sets using two-fold cross-validation and resulting probabilities of occurrence are used to predict vegetation type distributions within the study area. Predicted vegetation types are compared with observations, and the McNemar test indicates an overall better performance of the random forest model at the 0.001 significance level. Comparison of the modelling results for each individual vegetation type separately by means of the *F*-measure, which combines precision and recall, also reveals better predictions by the random forest model. Inspection of the probabilities of occurrence of the different vegetation types for each grid cell demonstrates that correct predictions in central areas of homogeneous vegetation sites are based on high probabilities, whereas the confidence decreases towards the margins of these areas. Threshold-independent evaluation of the model accuracy by means of the area under the receiver operating characteristic (ROC) curves confirms good performances of both models, but with higher values for the random forest model. Therefore, the incorporation of the random forest technique in distribution models has the ability to lead to better model performances.

© 2007 Elsevier B.V. All rights reserved.

* *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 © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.05.011

1. Introduction

Over the last decades, wetlands have lost their reputation of worthless land, and are now recognized as valuable areas (Mitsch and Gosselink, 2000a). Wetlands fulfil ecological, economical, protective and recreational functions, of which biodiversity conservation, water quality enhancement, and flood control are some examples (e.g. Gosselink and Turner, 1978; Kadlec and Knight, 1996; Mitsch and Gosselink, 2000b). Among the enormous variety of wetland types (Wheeler, 1999), this study focuses on lowland river ecosystems. They harbour a large part of biodiversity of the western European lowlands because of small-scale habitat diversification as a consequence of micro-topography, soil differences and differences in water sources, *i.e.* atmospheric water, groundwater, and river water (Wassen and Barendregt, 1992; De Becker et al., 1999; Bio et al., 2002). However, considerable losses of biodiversity in these valley ecosystems occur, mainly caused by high levels of N-deposition, groundwater abstraction, prevention of river flooding, agricultural drainage, application of fertilizers, and pollution of groundwater and surface water by sewage (Décamps et al., 1988; Schot and Molenaar, 1992; Erisman and Draaijers, 1992; Hellberg, 1995; Runhaar et al., 1996; all cited in Bio et al., 2002).

Ecohydrology tries to describe the hydrological mechanisms (like water availability and quality) that underlie ecological patterns and processes (Rodriguez-Iturbe, 2000). Within this scientific discipline, 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 (Venterink and Wassen, 1997; Ertsen et al., 1998). The models presented by Venterink and Wassen (1997) differ in scale level, habitat and ecosystem for which prediction was made, number of input variables, and expert knowledge and field measurements required. However, the relationship between response variable (e.g. the occurrence of species or vegetation types) and one or more explanatory variables (e.g. water table depth and water quality variables) was generally specified by a regression model (Bio et al., 1998). Ordinary multiple regression models and multiple logistic regression models within the frameworks of generalized linear models (GLM; McCullagh and Nelder, 1999) and generalized additive models (GAM; Hastie and Tibshirani, 1990; Yee and Mitchell, 1991) are very popular and are often used for modelling species distributions (Guisan and Zimmerman, 2000; Augustin et al., 2001; Austin, 2002; Engler et al., 2004; Rushton et al., 2004; Segurado and Araujo, 2004). Other predictive distribution models include neural networks (e.g. Fitzgerald, 1992; Recknagel, 2001); ordination (e.g. canonical correspondence analysis CCA; Hill, 1973) and classification methods (e.g. classification and regression trees; Breiman et al., 1984), Bayesian models (e.g. Fischer, 1990), environmental envelopes (e.g. Box, 1992) or even combinations of these models (Guisan and Zimmerman, 2000).

The objective of this study was to evaluate two different statistical techniques in a predictive ecohydrological modelling context. Therefore, a spatially distributed ecohydrological data set was used where 14 independent variables, describing the abiotic environment, were related with a binomial response variable, *i.e.* the occurrence (presence/absence) of vegetation types. The modelling techniques used were (i) multiple logistic regression (MLR) and (ii) random forest (RF). Multiple logistic regression in the generalized linear modelling framework has proven its applicability in ecological modelling in various studies (Guisan and Zimmerman, 2000; Austin, 2002; Stephenson et al., 2006). In GLMs, predictive variables are related to the response variable through a linear link function. Because of the binomial nature of the response variable of the data set a multiple logistic regression model (MLR) was used to describe the relationship between a combination of environmental predictive variables and the response variable. The second technique, random forest, is an ensemble learning technique, developed by Breiman (2001). Ensemble learning techniques generate many classifiers and aggregate their results (Liaw and Wiener, 2002). A random forest consists of a compilation of classification or regression trees (e.g. 1000 trees in a single random forest), and is empirically proven to be better than its individual members (Hamza and Larocque, 2005). This study only focuses on classification trees, since the response variable of the data set under investigation was the occurrence (presence/absence) of vegetation types. As with ordinary classification trees, each tree of the random forest assigns a class (here a vegetation type) to each measurement vector of environmental predictive variables. A majority vote over all trees in the forest defines the resulting response class.

2. Material and methods

2.1. Study sites

2.1.1. Data collection and site description

A data set collected from four sites (Zwarte Beek, Vorsdonkbos, Doode Bemde and Snoekengracht, Fig. 1) in different lowland river valleys in Flanders, the northern part of Belgium, was applied in this study (see also Bio et al., 2002). All sites are nature reserves with relatively undisturbed and unspoiled abiotic and biotic conditions, a long period of constant management (at least 10 years), and marked hydrological gradients. Data was collected in the period 1993–1997.

The Zwarte Beek site 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

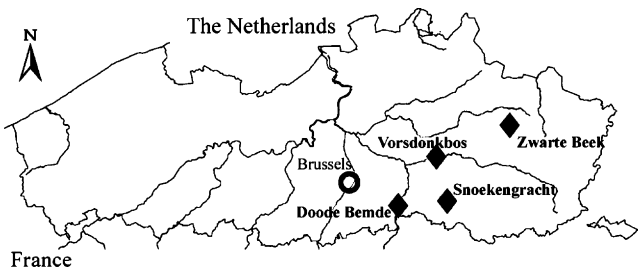


Fig. 1 – Location of the study sites in Flanders (northern Belgium).

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 and mineral-poor seepage water (ca. 16 mm day^{-1}). The groundwater table is constant and close to the surface level throughout the year (De Becker and Huybrechts, 2000a).

The Vorsdonkbos site 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 mineral-poor groundwater (20 mm day^{-1}). 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, which is fed by mineral-rich groundwater, the vegetation changes to tall herb fen (*Filipendulion*) and mesotrophic alder carr (*Caricion elongatae–Alnetum glutinosae*) (Huybrechts and De Becker, 2000).

The 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 mineral-rich groundwater (approximately 3 mm day^{-1}). 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* grasslands on the natural levees of the river (De Becker and Huybrechts, 2000b).

The Snoekengracht, situated approximately 57 m above sea level, is similar to the Doode Bemde site, except for a narrower valley and even more mineral-rich seepage water feeding the area (Huybrechts and De Becker, 1999).

2.1.2. *Abiotic site characterization*

The study sites were subdivided in regular 20 m ×20 m grid cells (10 m × 10 m grid cells for Snoekengracht). Soil type was derived from hand drillings at grid cell intersections to a depth of 1 m, classified using a set of four major texture types: sand, loam, clay and peat, and assigned to the neighbouring grid cells. Management regime 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 are 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. Groundwater level and quality was determined from samples collected from a piezometer network. Groundwater level was described by one variable: average groundwater depth (m). Groundwater depth samples were taken every 2 weeks in a 2-year period between 1993 and 1997. Groundwater quality variables were determined from groundwater samples taken during four different sampling campaigns in spring and autumn over two consecutive years within the period 1993–1997 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 = $100[1/2Ca^{2+}]/[1/2Ca^{2+} + Cl^-]$).

2.1.3. *Biotic site characterization*

During spring and early summer, in the period 1993–1997, plant species were mapped in the study sites on the same regular grid as soil texture and management regime. Mapping was restricted to about 85 mainly groundwater-dependent species. For each grid cell, species dominances and abundances were estimated using a decimal scale (Londo, 1976). Species cover data were used to define vegetation types for all study sites separately using TWINSPAN (Hill, 1979). Eleven clearly defined vegetation types were retained of which a short description is given in Table 1. Their spatial distribution is demonstrated in Fig. 2.

2.2. *The ecohydrological data set*

The groundwater quality variables measured at the piezometer point locations were spatially interpolated using block kriging (for details, see Bio et al., 2002; Huybrechts et al., 2002) in order to obtain groundwater variable estimates for all 1705 grid cells. Together with the other abiotic and biotic variables, groundwater quality 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 numerical and 2 categorical), describing the abiotic environmental conditions. Eleven different vegetation types $c_1, \ldots, c_{11}$ are considered (Table 1). 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):

$$L = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}. \quad (1)$$

2.3. *Statistical model description*

To meet the objective, *i.e.* to evaluate the random forest technique in a predictive ecohydrological modelling context, a widely used statistical modelling technique was selected for comparison. The choice for the multiple logistic regression technique was based on the binomial nature of the response variable in the ecohydrological data set, which is appropriate for analysis with this technique (Guisan and Zimmerman, 2000). Furthermore, the technique was used in earlier modelling studies for valley ecosystems in Flanders (Bio et al., 2002; Huybrechts et al., 2002) on the same ecohydrological data set.

2.3.1. *Multiple logistic regression*

Multiple logistic regression describes the relationship between a combination of environmental predictive variables and a binary response variable by means of a link function (Hosmer and Lemeshow, 2000). Consider a collection of p independent predictive variables denoted by the vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. Let the conditional probability that the outcome is 'present' be denoted by $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$ and the link function by $g(\mathbf{x})$, then (Hosmer and Lemeshow, 2000):

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \quad (2)$$

The logit$[\pi(\mathbf{x})]$ is used as link function in multiple logistic regression because of the binomial nature of the response variable. The link function $g(\mathbf{x})$ is given by

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \text{logit}[\pi(\mathbf{x})]. \quad (3)$$

Table 1 – Summary of the vegetation types: number, name, short description and area

No.	Name	Short description	Area (ha) (number of grid cells)			
			ZB 6.80 (170)	VB 12.80 (320)	DB 20.76 (519)	SG 6.69 (696)
1	*Alno–Padion*	Drier 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; DB, Doode Bemde; SG, Snoekengracht.

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 variable has k possible values, $k-1$ design variables are needed. When, for example, the jth predictive variable is soil type with four possible texture classes sand, loam, clay or peat, three design variables are necessary.

The link function for a GLM with p environmental predictive variables and the jth predictive variable being categorical would be

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \left(\sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} \right) + \cdots + \beta_p x_p = \text{logit}[\pi(\mathbf{x})], \tag{4}$$

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 vegetation type separately, in order to get an estimation of the probability of occurrence $\hat{\pi}(\mathbf{x})$ according to Eq. (2). Multiple logistic regression models were built using the S-plus statistical software. A full model, including first-order terms and quadratic variable terms (not included in Eqs. (3) and (4)), was fitted to the data using the likelihood function. Afterwards, stepwise insertion or deletion of variables was applied (Hosmer and Lemeshow, 2000). 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, Akaike, 1974), resulting in the model with the lowest AIC value.

2.3.2. *Random forests*

The random forest technique (Breiman, 2001) is an ensemble learning technique which generates many classification trees (Breiman et al., 1984) 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 individ-

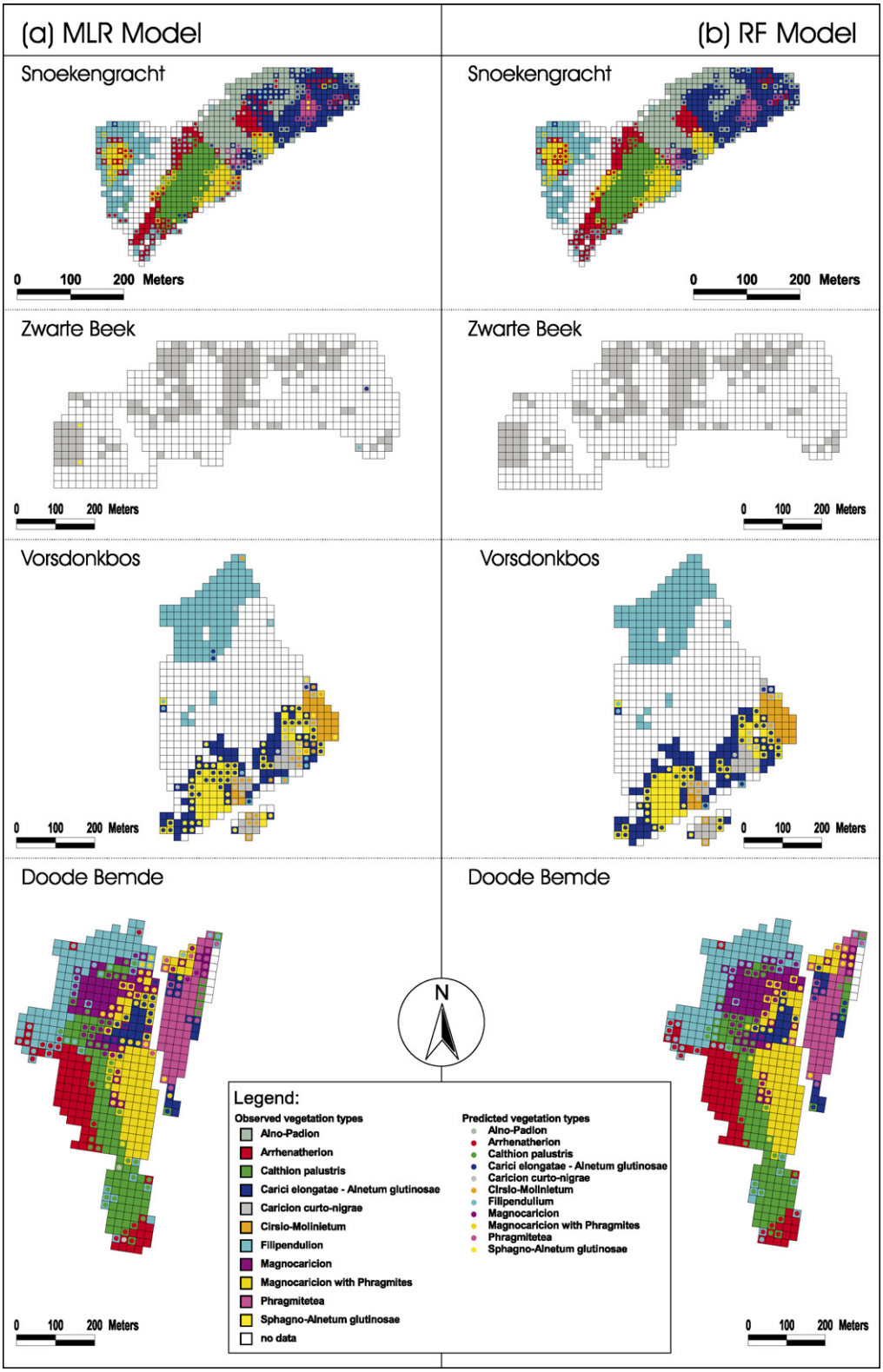


Fig. 2 – Predicted vegetation types with the multiple logistic regression model (a) and with the random forest model (b). The observed vegetation distribution (□) is overlaid with the predicted vegetation distribution (○). For each grid cell, the vegetation type with the highest probability of occurrence, as modelled with the multiple logistic regression model (a) and with the random forest model (b), is the predicted vegetation type.

ual members, is that the members of the ensemble perform better than random and are diverse (Hansen and Salamon, 1990). Random forests increase 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. 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). This is in contrast with standard classification tree building, where each node is split using the best split among all predictive variables. The algorithm for growing a random forest of k classification trees goes as follows:

(i) for $i = 1$ to k do:
  (1) draw a bootstrap subset X_i containing approximately 2/3 of the elements of the original data set X;
  (2) use X_i to grow an unpruned classification tree to the maximum depth, with the following modification compared with standard classification tree building: at each node, rather than choosing the best split among all predictive variables, randomly select m predictive variables and choose the best split among these variables;
(ii) predict new data according to the majority vote of the ensemble of k trees.

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. Predictive variables may be numerical or categorical, a translation to design variables is not needed.

An unbiased estimate of the generalization error is obtained during the construction of a random forest by

(i) for $i = 1$ to k do:
  (1) each 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 ith tree;
  (2) these out-of-bag elements are classified by the finalized ith tree.
(ii) At the end of the run, on average each element of the original data set X is out-of-bag in one-third of the k tree constructing iterations. Or, each element of the original data set is classified by one-third of the k trees. The proportion of misclassifications (%) over all out-of-bag elements is called the out-of-bag (oob) error.

The oob error is an unbiased estimate of the generalization error. Breiman (2001) 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 (Breiman, 2001): (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.

Some additional information generated by random forests is useful for ecohydrological modelling. The random forest technique estimates the importance of a predictive variable by looking at how much the oob error increases when oob data for that variable are permuted while all other variables are left unchanged. The increase in oob error is proportional to the predictive variable importance. Another measure assesses the proximity of different data points to one another. An $N \times N$ (with N the number of data points) proximity matrix is generated, with each element representing the fraction of trees in which the two corresponding data points fall in the same terminal node. The intuition is that similar data points should be in the same terminal node more often than dissimilar ones. Other measures and analysing options include variable interaction, missing value replacement and unsupervised learning (see Breiman and Cutler (2004a)). The use of these features, however, is beyond the scope of this study.

For random forest model development, Random Forests Version 5.1 (Breiman and Cutler, 2004b) was used. The randomForest package within the statistical software R 2.2.1 can also be used.

2.4. *Training versus test data sets*

The lack of an independent data set for model evaluation forced us to randomly and uniformly split the ecohydrological data set L into two parts. In two-fold cross-validation each of the two parts is once used as training set and once as test set:

$$L_{\text{train1}} = L_{\text{test2}} \text{ of size } 853, \tag{5}$$

$$L_{\text{train2}} = L_{\text{test1}} \text{ of size } 852. \tag{6}$$

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.

3. Model construction, calibration and results

3.1. *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. Each of these models consisted of 11 submodels, *i.e.* logit link functions $\hat{g}(\mathbf{x})$, one for each vegetation type. The submodels were constructed separately using the S-plus software in two steps: (i) submodel construction using all 14 variables as first-order

Table 2 – Model goodness-of-fit

Vegetation type	D_{null}	d.f.	D_{resid}	d.f.	$G = D_{null} - D_{resid}$	d.f.	Pearson resid.	d.f.
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 (constant only model); d.f., degrees of freedom; D_{resid}, residual deviance; G, the likelihood ratio test; Pearson resid., Pearson residuals. Significance at the 0.01 level (*) are indicated.

terms, and average groundwater depth, pH and $Fe_{(tot)}$ (which were identified as important predictive variables) as quadratic model terms, and (ii) bi-directional model term selection in a stepwise fashion using the AIC criterion. Casewise Pearson residual values (Hosmer and Lemeshow, 2000, p. 145) 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 ($L'_{train1} = 811$ training elements, and $L'_{train2} = 812$ training elements). Indications on model goodness-of-fit are given in Table 2. Null model deviances (constant only model), residual deviances, likelihood ratio test G and Pearson deviances are tabulated for MLR1 and MLR2. Since the deviance approximately follows a χ^2-distribution, this distribution is used to test upon. The residual deviances were all smaller than the corresponding χ^2-value at the 0.01 significance level. Therefore, the multiple logistic regression models were concluded to fit satisfactory (Neter et al., 1996). The same conclusion could be drawn based on the Pearson residuals. They proved a significant fit between observations and fitted values. The likelihood ratio test statistic G indicated that the multiple logistic regression models including significant predictive variables (as determined by the AIC criterion) fitted the observed vegetation type distribution better than the constant only models at the 0.01 significance level.

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 11 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 11 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 11 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. 2(a)). Out of the 1705 grid cells, 1182 (69.3%) were predicted correct, 524 (30.7%) incorrect. Despite its weaknesses (Vaughan and Ormerod, 2005), the Cohen's κ test (Cohen, 1960) was used to evaluate differences between observations and predictions. 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: there is a substantial agreement between observations and predictions ($p < 0.001$). Visual inspection of the results 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). However, for small and isolated patches and for boundary grid cells between neighbouring vegetation types, predictions were less accurate.

3.2. *Random forest model*

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.

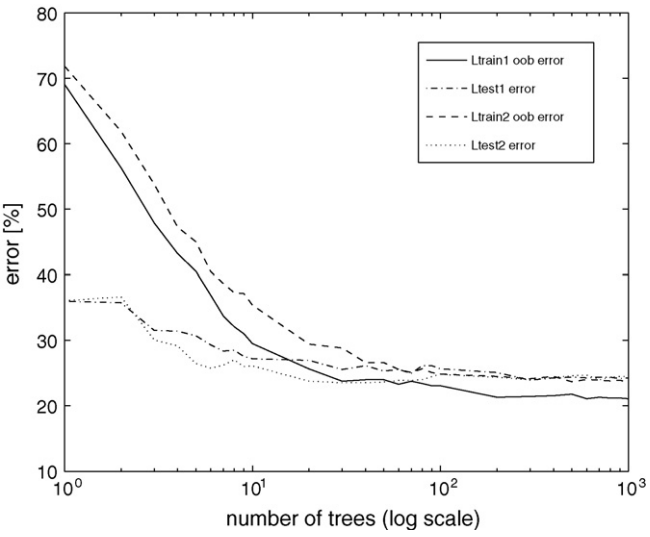


Fig. 3 – Out-of-bag (oob) error and test set error converge when more trees are added to the random forest. Ltrain1 oob error and Ltrain2 oob error are the oob errors calculated during the construction of RF1 and RF2, respectively. Ltest1 error and Ltest2 error are the test set error of RF1 and RF2 applied to their respective test data sets.

Breiman (2001) 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 1000 trees were constructed on L_{train1} and L_{train2}, respectively, both with two randomly selected variables to split the nodes ($m = 2$). Fig. 3 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 (Breiman, 2001), 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 on. The conclusions that could be drawn from Fig. 3 are (i) the oob error is a suitable estimator to detect error convergence, (ii) in accordance with Breiman (2001) 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 3. Both the oob error for RF1 and RF2 constructed on L_{train1} and L_{train2}, respectively, and 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 (2004a)). 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}.

Table 3 – 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	2	3	4	5	8	11	14
RF1[a]	21.78	20.26	19.91	20.37	20.61	20.02	20.37	21.19
RF2[a]	26.73	24.62	24.38	24.85	24.62	24.38	24.62	24.38
L_{test1} [b]	24.62	23.33	23.33	23.33	22.74	23.68	24.38	25.32
L_{test2} [b]	25.06	23.77	23.42	23.77	24.24	24.36	24.71	25.17

Minimal values are in italics.
[a] Oob error.
[b] Test set error.

Each measurement vector $\mathbf{x}_i$ of the test sets was classified by each tree as a unique vegetation type. 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:

$$P(c_j) = N_{c_j}/N_{tot}, \quad (7)$$

where $P(c_j)$ 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} the total number of classification trees in the random forest (here $N_{tot} = 1000$).

This probability of occurrence was calculated for the 11 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. 2(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. A κ value of 0.734 was found: 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 (0.651).

4. Model evaluation

4.1. *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 two-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.

4.1.1. *McNemar test*

For $L = L_{test1} \cup L_{test2}$ (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 of Dietterich (1998), the McNemar test (Everitt, 1992) 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. 2) were compared with the observations and used to construct the following contingency table:

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}

where $N = n_{00} + n_{01} + n_{10} + n_{11}$ is the total number of elements in the ecohydrological data set. 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:

$$M = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}. \quad (8)$$

If the null hypothesis is correct, then the probability that this quantity is greater than $\chi^2_{1,0.95} = 3.84$ 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.

4.1.2. *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 11 vegetation types separately using a confusion matrix (e.g. Fielding and Bell, 1997; Kohavi and Provost, 1998). Several standard terms have been defined for a confusion matrix (Fielding and Bell, 1997; Kohavi and Provost, 1998) 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' (Van Rijsbergen, 1979). A weighted version of the F-measure was used:

$$F_\beta(p, r) = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}, \quad (9)$$

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 almost no correct predictions are made, 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 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 (Wilcoxon, 1945) 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* 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. 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* and *Alno–Padion*– were predicted to be *Spagno–Alnetum glutinosae* and many cells – mainly *Magnocaricion* with *Phragmites* and *Calthion palustris*– were predicted to be *Magnocaricion* (see Fig. 2). This is somewhat understandable as these are spatially adjacent, comparable vegetation types with dominance of *Alnus glutinosa* (L.) Gaertn. in *Sphagno–Alnetum glutinosae*, *Alno–Padion* and *Carici elongetae–Alnetum glutinosae*, and the higher abundance of *Phragmites australis* as main difference between *Magnocaricion* and *Magnocaricion* with *Phragmites*. Recall was lowest for *Sphagno–Alnetum glutinosae* and *Magnocaricion* for the multiple logistic regression and the random forest model. In Fig. 2 the large number of wrong predictions for *Sphagno–Alnetum glutinosae* and *Magnocaricion* in Vorsdonkbos 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 (marked lower Mg^{2+}, Ca^{2+} and Cl^- concentrations).

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 11 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 ($p = 0.003$).

4.2. *Prediction probabilities*

4.2.1. *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). 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 towards 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.

4.2.2. *Threshold-independent evaluation*

Receiver operating characteristic (ROC) curves are frequently used for the evaluation of classification accuracy. This curve, originating from signal detection theory, is widely used in clinical sciences, but recently also in earth sciences (Guisan and Zimmerman, 2000; Pontius and Schneider, 2001; Boyce et al., 2002; Liu et al., 2005; Phillips et al., 2006). In ROC space, one plots the false positive rate (FPR) on the *x*-axis and the true positive rate (TPR, =*recall*) on the *y*-axis. The FPR (FP/(FP +TN)) measures the fraction of negative grid cells (*i.e.* vegetation type absent) that are incorrectly predicted as positive (*i.e.* vegetation type present). The TPR (TP/(TP +FN)) measures the fraction of positive grid cells that are predicted correctly.

Table 4 – McNemar test; precision, recall, $F_{0.5}$, F_1, F_2 for MLR and RF

	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. (8).

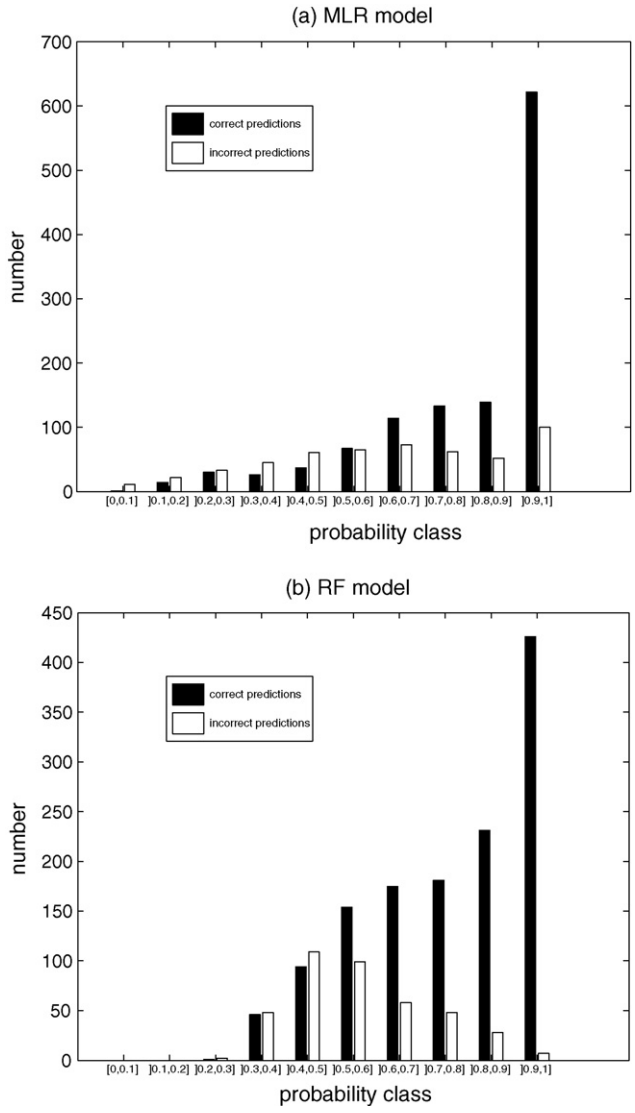


Fig. 4 – Probability distributions of predictions made with the multiple logistic regression model (a) and the random forest model (b).

The multiple logistic regression model and the random forest model computed the probabilities of occurrence of 11 vegetation types. Earlier we used the decision rule that the most probable vegetation type (among the 11 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. 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 (Hosmer and Lemeshow, 2000). 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. Both models had high AUC-values, reflecting their excellent discrimination abilities (Table 5). *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. The Wilcoxon signed rank statistic indicated significantly higher median AUC-values for the random forest model at the 0.01 significance level.

Table 5 – 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.

5. Discussion and conclusions

5.1. *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 11 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 two-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).

(3) The *F*-measures, which combines precision and recall, were significantly better for the random forest model.
(4) The multiple logistic regression model made correct predictions with higher probabilities than the random forest model (Fig. 4). 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 study is that the random forest modelling technique has the ability to lead to better predictive ecohydrological models.

5.2. *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. Diaz-Uriate and de Andres, 2006; Chen and Liu, 2006) and within the earth-sciences in remote sensing (e.g. Pal, 2005; Ham et al., 2005; Gislason et al., 2006). However, no example of the use of the random forest technique in ecological distribution modelling was found, and therefore comparison possibilities with literature were few. 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 predictive variables. Austin (2002) 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 (Austin, 2002). The most proximal gradient will be the causal variable determining the plant response. When proximal resources and direct gradients are used as environmental predictive 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 (Guisan and Thuiller, 2005) because 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 (Franklin, 2000).

In order to gain functionality of the random forest model, further research should focus on its modelling ability with smaller data subsets, comprising uncorrelated (most likely) proximal predictive variables. There are several reasons to do so (Mac Nally, 2000): (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 two-fold cross-validation data and on spatially independent evaluation data tend to differ (Edwards et al., 2006).

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 grateful to the Research Programme on Nature Development (projects VLINA 96/03 and VLINA 00/16) of the Flemish Government. The first author wishes to thank Inge De Jongh and Rudi Hoeben for computer assistance, and Heidi Wouters for statistical help. The authors are grateful to the reviewers for their constructive comments.

REFERENCES

Akaike, H., 1974. A new look at statistical-model identification. IEEE Trans. Autom. Contr. 19 (6), 716–723.
Augustin, N.H., Cummins, R.P., French, D.D., 2001. Exploring spatial vegetation dynamics using logistic regression and a multinomial logit model. J. Appl. Ecol. 38, 991–1006.
Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157 (2/3), 101–118.
Bio, A.M.F., Alkemade, R., Barendregt, A., 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. J. Veget. Sci. 9, 5–16.
Bio, A.M.F., De Becker, P., De Bie, E., Huybrechts, W., Wassen, M., 2002. Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions. Biodivers. Conserv. 11, 2189–2216.
Box, E.O., 1992. Macroclimate and Plant Forms: An Introduction to Predictive Modeling in Phytogeography. Junk, The Hague.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157 (2/3), 281–300.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Cutler, A., 2004a. http://www.stat.berkeley.edu/users/Breiman/RandomForests/cc_papers.html.
Breiman, L., Cutler, A., 2004b. http://www.stat.berkeley.edu/users/Breiman/RandomForests/cc_software.html.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall, New York.
Chen, X.W., Liu, M., 2006. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21 (24), 4394–4400.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.
De Becker, P., Hermy, M., Butaye, J., 1999. Ecohydrological characterization of a groundwater-fed alluvial floodplane mire. Appl. Veget. Sci. 2, 215–228.
De Becker, P., Huybrechts, W., 2000a. Vallei van de Zwarte Beek—Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium (in Dutch).
De Becker, P., Huybrechts, W., 2000b. De Doode Bemde—Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium (in Dutch).
Décamps, H., Fortuné, M., Gazelle, F., Patou, G., 1988. Historical influence of man on riperian dynamics of a fluvial landscape. Landsc. Ecol. 1, 163–173.
Diaz-Uriate, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923.
Edwards, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen Jr., G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecol. Model. 199, 132–141.
Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.
Erisman, J.W., Draaijers, G.P.J., 1992. Atmospheric Deposition in Relation to Acidification and Eutrofication. No. 63 in Studies in Environmental Science. Elsevier, Amsterdam.
Ertsen, A.C.D., Bio, A.M.F., Bleuten, W., Wassen, M.J., 1998. Comparison of the performance of species response models in several landscape units in the province of North-Holland, The Netherlands. Ecol. Model. 109 (2), 213–223.
Everitt, B.S., 1992. The Analysis of Contingency Tables, 2nd ed. Chapman and Hall, London.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24 (1), 38–49.
Fischer, H.S., 1990. Simulating the distribution of plant communities in an alpine landscape. Coenoses 5, 37–49.
Fitzgerald, R.W., 1992. The application of neural networks to the floristic classification of remote sensing and gis data in complex terrain. In: Proceedings of the XVII Congress ASPRS, American Society of Photogrammetry and Remote Sensing, Bethesda, MD, pp. 570–573.
Franklin, J., 2000. Predicting the distribution of shrub in southern California from climate and terrain-derived variables. J. Veget. Sci. 9, 733–748.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recogn. Lett. 27 (4), 294–300.
Gosselink, J.G., Turner, R.E., 1978. The role of hydrology in freshwater wetland ecosystems. In: Good, R.E., Whingham, D.F., Simpson, R.L. (Eds.), Freshwater Wetlands: Ecological Processes and Management Potential. Academic Press, pp. 63–78.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.
Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2/3), 147–186.
Ham, J., Chen, Y.C., Crawford, M.P., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43 (3), 492–501.
Hamza, M., Larocque, D., 2005. An empirical comparison of ensemble methods based on classification trees. J. Statist. Comput. Simulat. 75 (8), 629–643.
Hansen, L., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12, 993–1001.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
Hellberg, F., 1995. Entwicklung der Grünland Vegetation bei Wiedervernässung und periodischer Überflutung. Vegetationsökologische Untersuchungen in nordwestdeutschen Überflutungspoldern. Vol. 243 of Dissertationes Botanicae. Balogh Scientific Books.
Hill, M.O., 1973. Reciprocal averaging—eigenvector method of ordination. J. Ecol. 61, 237–244.
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, New York.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. Wiley, Chichester.
Huybrechts, W., De Becker, P., 1999. De Snoekengracht—Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium (in Dutch).
Huybrechts, W., De Becker, P., 2000. Vorsdonkbos - Turfputten—Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium (in Dutch).
Huybrechts, W., De Becker, P., De Bie, E., Wassen, E., Bio, A., 2002. Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, ITORS-VL. VLINA 00/16. Institute of Nature Conservation, Brussels, Belgium (in Dutch).
Kadlec, R.H., Knight, R.L., 1996. Treatment Wetlands. Lewis Publishers, Boca Raton.
Kohavi, R., Provost, F., 1998. Glossary and terms. Mach. Learn. 30, 271–274.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2/3, 18–22.
Liu, C.R., Berrey, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28 (3), 385–393.
Londo, G., 1976. Decimal scale for releves of permanent quadrats. Vegetatio 33, 61–64.
Mac Nally, R., 2000. Regression and model-building in conservation biology, biogeography, and ecology: the distinction between – and reconciliation of – 'predictive' and 'explanatory' models. Biodivers. Conserv. 9, 655–671.
McCullagh, P., Nelder, J.A., 1999. Generalized Linear Models, 2nd ed. Chapman and Hall, Boca Raton.
Mitsch, W.J., Gosselink, J.G., 2000a. Wetlands, 3rd ed. Wiley & Sons, New York.
Mitsch, W.J., Gosselink, J.G., 2000b. The value of wetlands: importance of scale and landscape setting. Ecol. Econ. 35, 25–33.
Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models, 4th ed. WCB/McGraw-Hill, USA.

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26 (1), 217–222.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (3/4), 231–259.
Pontius, R.G., Schneider Jr., L., 2001. Land-use change model validation by a ROC method for the Ipswich watershed, Massachusetts, USA. Agric. Ecosyst. Environ. 85 (1–3), 239–248.
Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146, 303–310.
Rodriguez-Iturbe, I., 2000. Ecohydrology: a hydrologic perspective of climate–soil–vegetation dynamics. Water Resour. Res. 36, 3–9.
Runhaar, J., Vangool, C.R., Groen, C.L.G., 1996. Impact of hydrological changes on nature conservation areas in the Netherlands. Biol. Conserv. 76, 269–276.
Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modeling species distributions. J. Appl. Ecol. 41, 193–200.
Schot, P.P., Molenaar, A., 1992. Regional changes in groundwater-flow patterns and effects on groundwater composition. J. Hydrol. 130, 151–170.
Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modeling species distributions. J. Biogeogr. 31, 1555–1568.
Stephenson, C.M., MacKenzie, M., Edwards, C., Travis, J., 2006. Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. Ecol. Model. 193, 747–758.
Van Rijsbergen, C.J., 1979. Information Retrieval, 2nd ed. Butterworths, London.
Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution models. J. Appl. Ecol. 42, 720–730.
Venterink, H.O., Wassen, M.J., 1997. A comparison of six models predicting vegetation response to hydrological habitat change. Ecol. Model. 101 (2/3), 347–361.
Wassen, M.J., Barendregt, A., 1992. Topographic position and water chemistry of fens in a Dutch riverplain. J. Veget. Sci. 3, 447–456.
Wheeler, B.D., 1999. Water and plants in freshwater wetlands. In: Baird, A.J., Wilby, R.L. (Eds.), Eco-hydrology: Plants and Water in Terrestrial and Aquatic Environments. Routledge, London, pp. 127–180.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1, 80–83.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive-models in plant ecology. J. Veget. Sci. 2, 587–602.