Environmental Modelling & Software 92 (2017) 269–316

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador

Sacha Gobeyn [a, *], Martin Volk [b], Luis Dominguez-Granda [c], Peter L.M. Goethals [a]

[a] Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, B-9000 Ghent, Belgium
[b] UFZ - Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany
[c] Centro del Agua y Desarrollo Sustentable, Escuela Superior Politécnica del Litoral (ESPOL), Campus Gustavo Galindo, Km. 30.5 Via Perimetral, PO Box 09-01-5863, Guayaquil, Ecuador

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form
6 January 2017
Accepted 15 February 2017

Keywords:
Conceptual species distribution models
Input variable selection
Simple genetic algorithms
Species response curves
River pollution
Freshwater management

ABSTRACT

Species distribution models (SDMs) have received increasing attention in freshwater management to support decision making. Existing SDMs are mainly data-driven and often developed with statistical and machine learning methods but with little consideration of hypothetic ecological knowledge. Conceptual SDMs exist, but lack in performance, making them less interesting for decision management. Therefore, there is a need for model identification tools that search for alternative model formulations. This paper presents a methodology, illustrated with the example of river pollution in Ecuador, using a simple genetic algorithm (SGA) to identify well performing SDMs by means of an input variable selection (IVS). An analysis for 14 macroinvertebrate taxa shows that the SGA is able to identify well performing SDMs. It is observed that uncertainty on the model structure is relatively large. The developed tool can aid model developers and decision makers to obtain insights in driving factors shaping the species assemblage.

Software availability

Name: Species Distribution Model Identification Tool (SDMIT)
Developer: Sacha Gobeyn, Coupure Links 653, B-9000 Ghent, sacha.gobeyn@ugent.be, sachagobeyn@gmail.com
Program language: Python 2.7.10
Software requirement: Anaconda Python distribution package (Continuum Analytics, https://www.continuum.io/why-anaconda)
Source code: https://github.com/Sachagobeyn/SDMIT/releases/tag/v1.0.0, https://github.com/Sachagobeyn/SDMIT
License: CC BY 4.0 Creative Commons

1. Introduction

Species distribution models (SDMs) are increasingly used by scientists and policy makers as tools to investigate a wide variety of ecological problems. With respect to freshwater management, different SDMs are developed and used to quantify the impact of anthropogenic pressures (e.g. increased nutrient concentrations, loss of habitat) on the distribution and diversity of macroinvertebrate taxa (Van Broekhoven et al., 2006; Domisch et al., 2013; Boets et al., 2015; Gies et al., 2015). Statistical and machine learning methods, used to identify SDMs with data, are useful, but have also been criticised because they often lack in conceptual background of ecology theory (Austin, 2007; Fukuda et al., 2013).

All SDMs are positioned along an axis of purely data-driven (data influenced) to conceptual (hypothetic influenced) models (Beale and Lennon, 2012; Mount et al., 2016). No model is solely data-driven or conceptual, but current SDM approaches often rely on statistical and machine learning approaches rather than on hypothetic knowledge (Austin, 2007). This causes the SDMs being characterised by statistical assumptions and properties of the used ecological data. Consequently, the models are vulnerable to imperfections and uncertainties in the data.

In contrast, conceptual SDMs primarily aim to reflect ecological concepts. For instance, BIOCLIM, one of the first SDMs, aims to delineate a number of environmental envelopes reflecting the total

* Corresponding author.
E-mail address: Sacha.Gobeyn@UGent.be (S. Gobeyn).

http://dx.doi.org/10.1016/j.envsoft.2017.02.012
1364-8152/

range of environmental space permitting a positive growth of a species (Booth et al., 2014). This approach is directly linked with the concepts of niche theory presented by Hutchinson (1957). In another example, the Genetic Algorithm for Rule Set Production (GARP) modelling system of Stockwell and Noble (1992) is an approach explicitly searching for heuristic rules of environmental boundaries so to explain species presence. In this approach, environmental thresholds, which reflect the boundaries in the environmental space limiting species distribution, are searched. Although both methods know many applications, they are often outperformed by novel machine learning approaches, making them possibly less interesting for application (Elith et al., 2006; Hamblin, 2013).

The interface between data-driven and conceptual models poses the challenge to develop perfomant SDMs with sufficient conceptual ground (Austin, 2007; Guisan and Rahbek, 2011; Bennetsen et al., 2016). Starting with conceptual approaches, one is challenged to identify well performing models. Identifying (an) alternate model(s) from a set of models with data is often a difficult task, especially when many candidate models can be formulated (i.e. different possible input variables, interactions, model parameters). Computer algorithms, using metaheuristics, have proven valuable to tackle this challenge (Maier et al., 2014).

In this paper, an approach using a simple genetic algorithm (SGA), a type of heuristic search algorithm, is presented to identify alternative formulations for a conceptual SDM. The process of identifying alternative formulations is defined here as model identification, and consists of identifying values for the parameters, selection of input variables and interaction functions. The approach is illustrated by implementing and using an SGA to perform input variable selection (IVS), for a case study of ecological water quality (EWQ) assessment and river pollution in the Guayas River basin, Ecuador. The case study is selected because the anthropogenic pressures on the river system are relatively straightforward to interpret. The remainder of the paper is structured as follows: section 2 presents a methodological framework for model identification of conceptual models with SGAs and evolutionary algorithms (EAs) in general. In section 3, the methodology is illustrated with the Guayas River case study and in section 4, the approach and results are discussed.

2. Model identification in conceptual SDMs

In this section, we present an approach to identify alternative formulations for a conceptual SDM. In the first part (section 2.1), the concepts commonly used in species distribution modelling are shortly addressed. In section 2.2, an approach to develop a conceptual SDMs is presented, with specific focus on model identification. In the last part (section 2.3), the use of EAs as a tool to identify alternative models is presented. SGAs are classified under EAs and are used as specific implementation of EAs in section 3.

2.1. Preliminaries

The basis for every model is the definition of a conceptual model, which involves the formulation of assumptions and underlying ecological theories and processes (Guisan and Zimmermann, 2000). Filter theory is an attractive starting point for a conceptual model, because this theory attempts to structure the processes driving species absence in a number of elements (Poff, 1997; Guisan and Rahbek, 2011). In filter theory, it is assumed that the realized species assemblage is the results of hierarchical filters, functioning as limits or barriers for species to exist in a local community. Typically, three types of filters are distinguished, each (inter-)acting on a specific scale: the dispersal, abiotic and biotic filters (Guisan and Rahbek, 2011). These hierarchical filters are assumed to act upon species traits, which are functional attributes of the species (for instance, body size, weight, reproduction, .).

Conceptually, filter theory is closely related to niche theory of Grinnell (1917) and Hutchinson (1957). A niche is defined as the environmental space which permits an intrinsic growth rate of a species. In the note of Hutchinson (1957), a distinction is made between a fundamental and realized niche. A realized niche is defined as a subset of a fundamental niche, not only constrained by abiotic factors, but also by dispersal (and historical) limitations and biotic interactions between the species (Soberón and Nakamura, 2009). In species distribution modelling, it is assumed that a realized niche is fitted, because in reality, species are observed in this niche (Guisan and Thuiller, 2005). However, Beale and Lennon (2012) state that it is preferable to model a fundamental niche rather than a realized niche, because the narrower precision of a realized niche might underestimate model uncertainty. An insight in this uncertainty is of major importance in - for instance - estimating the effect of climate change on species, since simulations are run for environmental conditions outside the fitted range.

2.2. Model identification in species distribution modelling

In Fig. 1, an approach for the development of a conceptual SDM is presented. The first step is to define and use a number of theoretical ecological concepts for the basis of the model (see section 2.1). Typically, filter theory and/or niche theory are used as an initial basis, because of their structural approach, subdividing driving processes in a number of elements (see for instance Guisan and Rahbek (2011) and Poff (1997)). With respect to species distribution modelling, the concept of habitat suitability or species response curves, defining the biological response to abiotic conditions, is often used to reflect the concept of abiotic filters of filter theory. In a second step, information on the attributes of the system are collected from databases, field samples and experts. Next, the model is constructed by defining a number of structural elements; input and output variables, interactions and parameters with data (see Bennetsen et al. (2016)) and/or expert knowledge (see Adriaenssens et al. (2006)). In the final step, alternative model structures are identified, typically by testing the conceptual model with data. Machine learning and statistical approaches can be used as a means to explore the space of possible alternative models. The

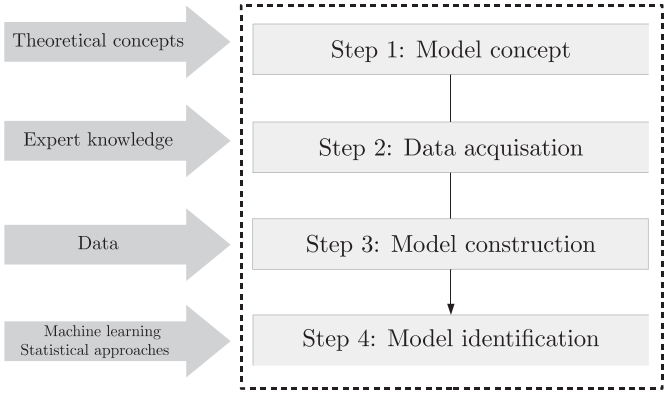

Fig. 1. Four step approach for model development of conceptual SDMs (adapted after Bennetsen et al. (2016)). In the first step, the used theoretical ecological concepts are formulated, whereas in the second step, information on the model attributes is collected. In step three, the model is constructed with the available information and finally, alternative model structures are identified. In all of these steps, ecological concepts, acquired information, expert knowledge and machine learning and statistical approaches can be used to justify and aid development.

dimensions of the search space can be reduced by using expert knowledge to define boundary conditions to the search (Kissling et al., 2012) and/or assuming model simplifications (Maier et al., 2014). It is important to note that in this four step approach, the influence of data and its accompanied uncertainties in the model increases. In order to cope with these uncertainties, one is advised to estimate the effect of data uncertainties on the SDM.

2.3. The use of evolutionary algorithms for model identification

Evolutionary algorithms (EAs) classified under metaheuristic methods, are a set of methods which incorporate elements of structured randomness for search, based on phenomena observed in nature (selection, mutation, crossover) (Goldberg, 1989; Maier et al., 2014). They aim to provide answers to complex questions, and more specifically applied to species distribution modelling and model identification, search for (near-)optimal solutions for IVS (D'heygere et al., 2003, 2006; Boets et al., 2013) and parameter estimation (Fukuda et al., 2012). In an EA, a candidate solution to a problem (i.e. phenotype) is represented by an object, also called a genotype. The data structure of a genotype, programmed in a chromosome, is often a binary string, a real-valued string or a tree (Eiben and Smith, 2015). EAs require a mapper function, transforming the genotype to phenotype, and an objective function, calculating the fitness of the phenotype. In model identification, the objective function is often calculated by running the candidate solution (i.e. candidate model) and testing the model output with data.

In this paper, we advocate the use of EAs for model identification of SDMs. One key advantage of EAs is that the flexibility in the way the data structure for the genotype is programmed provides the opportunity for users to relatively easily define their model identification problem. Users can choose to implement any sort of data structure, as long as the structure can be translated by a genotype-phenotype mappper. This allows to code IVS (e.g. 0 = absent, 1 = present), parameter estimation (e.g. binary representation of real-values) and selection of aggregation (i.e. interaction) functions of SDMs in a relative straightforward way, without spending time on implementing the operators of the search algorithm.

3. Experimental study

IVS is, besides parameter estimation and selection of aggregation functions, an essential component in the model identification process, since the input variables contain the information on the main factors which define the system under research. We aim to illustrate IVS with an SGA for a case study of water pollution in the Guayas River. To do so, we develop SDMs with abiotic filtering for 14 macroinvertebrate taxa (see Table 1). Abiotic and biological data collected in the Guayas River basin are used (section 3.2). Since the focus of the case is water pollution, we implemented models with abiotic filters describing the state of pollution. Expert knowledge was considered for this case study, however, no expert knowledge database was found to provide as a basis for the construction of the models (see approach Adriaenssens et al. (2006)). Since environmental conditions between South-America and other continents may differ significantly, it is preferable to not transfer the expert knowledge from other continental ecological knowledge databases. The taxa are selected based on their sensitivity to water pollution and sample prevalence (presence over total number of samples). The sample prevalence is calculated with the data described in section 3.2. The tolerance score (TS) of the Biological Monitoring Working Party for Colombia (BMWP-C) index, which is an adapted index that gives a measure of the EWQ using macroinvertebrates families as biological indicators, is used as a measure of sensitivity to pollution (Table 1).

Table 1
Overview of taxa used in the study. The taxa are selected based on their sample prevalence (Pr, %, calculated with the data described in section 3.2) and tolerance score (-, measure for sensitivity, $\in [0,10]$, 0 = not sensitive, 10 = sensitive). Note that prevalence is equal to sample prevalence (not species prevalence) and does not explicitly link to the rarity of the species.

Taxon	TS (−)	Pr (%)	Taxon	TS (−)	Pr (%)
Chironomidae	2	83	Hydrophilidae	3	11
Acari	0	47	Tubificidae	1	24
Baetidae	7	53	Dytiscidae	9	11
Libellulidae	6	46	Gomphidae	10	14
Leptohyphidae	7	44	Caenidae	7	10
Coenagrionidae	7	42	Hydroptilidae	7	10
Thiaridae	5	30	Corydalidae	6	9

3.1. Conceptual model

In this paper, the case study of water pollution is investigated as an illustration of the methodology presented in section 2. Freshwater macroinvertebrates are investigated since their response to water pollution is considered as a good indicator for pollution (Gabriels et al., 2010). The focus in this paper will be pollution, and that is why we consider abiotic filtering in this study.

Species response curves (or habitat preference curves) defining the biological responses to abiotic gradients are used as the basis for the abiotic filters. The biological response can be expressed by many measures, e.g. species presence, abundance, density, usable area or volume. Species abundance and presence are considered as good measures for biological response (Fukuda et al., 2012). Species abundance is not used often in species distribution modelling, but is investigated as a potential source of information to improve SDMs (Fukuda et al., 2012; Howard et al., 2014). Considering abundance data may help to provide detailed information for the SDMs, however, the uncertainties inherit to the sampling design and errors of abundance observations can cause to degrade the precision of the SDMs. In contrast, presence data contain simple information, which is assumed to be less uncertain (Fukuda et al., 2012; Gobeyn et al., 2016). Hence, we used species presence as a measure to express biological response.

The broadest range of environmental conditions for which the taxa are observed is used to define the species response curves. The aim of this approach is to recover a fundamental niche, however, it is highly likely that an approximation between a realized and a fundamental niche is obtained. This will increase model uncertainties, but is considered more realistic when predicting future distributions (Beale and Lennon, 2012). The shape of the species response curve is a much discussed topic, to which no agreement has been reached yet (Austin, 2002). The commonly used unimodal and symmetric Gaussian response curves are usually not in conjunction with empirical data, which suggests non-symmetrical responses are also possible, especially in case of species interactions and extreme environmental stress (Austin, 2007; Hirzel and Le Lay, 2008; Heikkinen and Mäkipää, 2010). We used a non-symmetric unimodal trapezoid curve as a simplification of the bell-shaped curve (Guisan and Zimmermann, 2000; Austin, 2002, 2007). The shape of the curve is defined by four parameters, a_1, a_2, a_3 and a_4. By defining the trapezoid curves by four parameters, one allows to describe a simplified form of various types of distributions with different degrees of skewness and kurtosis. This allows to test multiple shapes of response curves, and does not require the apriori definition of a shape of response. This approach is used, since in literature, no definite proof is provided for one or the other shape of response (Austin, 2007).

3.2. Study area and database

The Guayas River basin is located in central-western Ecuador. With 33700 km^2 land surface area, it is the largest watershed in South America west of the Andes mountains. Flows of the Guayas River can reach up to 5000 m^3/s (Waite, 1982). It is considered the most important river in Ecuador due to the industrial and economical value. Since the river basin is used to provide cultivated foods and water supply for human use, it is of major importance to mitigate effects of anthropogenic pressures on the ecosystem's quality and functioning (Arias-Hidalgo et al., 2013). A descriptive analysis of the EWQ is presented by Ambarita et al. (2016). The authors used a correspondence analysis (CA) to identify the link between different habitat characteristics and the BMWP-C. They concluded that upstream sites have a better EWQ than the sites downstream. In addition, flow velocity, chlorophyll concentration, conductivity, land use, sludge layer and sediment type are identified as important environmental variables determining the EWQ. It is hypothesized that nutrients and pesticides, coming from agriculture, and poorly treated waste water affect the EWQ. In our study, we aim to test whether a relation can be identified between some key species and nutrient concentrations.

Samples of the macroinvertebrate community were collected at 120 sites in the Guayas River basin during the dry season of 2013 (October to November). All sites were sampled once with the kick sampling procedure described by De Pauw and Vanhooren (1983) and Gabriels et al. (2010). A handnet, with a frame size of 20 × 30 cm and a mesh size of 500 μm, was used to sample during 5 minutes a 10–20 m stretch of the river. The sampling efforts were distributed over all aquatic habitats, including bed substrates, macrophytes, artificial and other (e.g. floating, submerged) substrates. In a laboratory, the handnet samples are rinsed with a 500 μm sieve and distributed over several trays. In these trays, macroinvertebrates are picked and after collection identified at family level with the identification key of De Pauw and Vannevel (1991) and Domínguez and Fernández (2009). The abundances of all present taxa were counted or estimated in the tray. For each sample, multiple physico-chemical water quality characteristics were measured in situ using two YSI6920-V2 multiparameter probes; temperature (°C), conductivity (μS cm^{-1}), total dissolved solids (TDS, g L^{-1}), pH (–), chlorophyll (mg L^{-1}), chloride (Cl$^-$, mg L^{-1}), dissolved oxygen (DO, mg O_2 L^{-1} and %) and turbidity (NTU). Concentrations of chemical oxygen demand (COD, mg L^{-1}), total nitrogen (total N, mg L^{-1}), total phosphorus (total P, mg L^{-1}), nitrate-N (NO_3^+-N, mg L), nitrite-N (NO_2-N, mg L^{-1}) and ammonium-N (NH_4^+-N, mg L^{-1}) were measured by taking water samples, stored cool and dark, and analysing these in a laboratory using HachLangeDR 3900 spectrophotometer kits (320–1100 nm wavelength range and 1 nm wavelength resolution). Additionally, sampling site elevation was measured with a Garmin GPSMap (Forio et al., 2015; Ambarita et al., 2016).

The coupled abiotic-biological data are processed (see Appendix A) and sampled so a subset can be used for the construction of the model (section 3.3) and the other subset for model identification (secion 3.4). Here, we refer to a subset of the data as a data sample. A data sample is obtained by resampling the data according to four criteria. Here, we aim to divide the data in two subsets, used for model construction and identification. In this approach, the data are selected so that the two samples have an equal number of instances for system type (lake and river) and EWQ classes (based on BMWP-C). The distance between measurement points is chosen as a last sample criterion to ensure a homogeneous spatial coverage in the data sample. A last criteria required that the same ratio of presence to absence is present in both sets.

3.3. Model construction

Fig. 2 illustrates the approach to estimate the species response curves. The species response curve of the taxon Y for an abiotic gradient X (SI_Y (X)) is estimated by inspecting the range of abiotic conditions for which a taxon is observed. The minimum, 25, 75 percentile and maximum of the empirical distribution for the studied gradient are used to define the optimal range (a_2 and a_3) and total range (a_1 and a_4) of the response (see Fig. 2). Note that the definition of these curves is closely related to the definition of envelope models, where the response to a gradient is described by a(n) (rectangular) environmental envelope (Carpenter et al., 1993). In addition, it is important to note that the choice for a_2 and a_3 (respectively 25 and 75 percentile) to define the optimal environmental range (X:SI_Y (X) = 1) is an arbitrary choice. The selection of a_2 and a_3 presents a trade-off between robustness and the optimal range that is described. In other words, lower and higher percentiles for respectively a_2 and a_3 can be considered to describe the optimal range. This will result in a broader but less robust range. Other percentiles are considered, however, the 25 and 75 percentiles are assessed as the most robust describing a relative wide optimal range.

The suitability index (SI) is calculated by applying a species response curve (Fig. 2) for an input value x_k^j:

$$\mathrm{SI}\left(x_k^j\right) = \begin{cases} 0 & \text{if } x_k^j < a_1 \\ \dfrac{x_k^j - a_1}{a_2 - a_1} & \text{if } x_k^j \in [a_1, a_2] \\ 1 & \text{if } x_k^j \in [a_2, a_3] \\ \dfrac{a_4 - x_k^j}{a_4 - a_3} & \text{if } x_k^j \in [a_3, a_4] \\ 0 & \text{if } a_4 < x_k^j \end{cases} \quad (1)$$

j is the index of the variable (m variables) and k, the index of the data point (n data points). The suitability indices are aggregated to one value, the habitat suitability index (HSI) (Raleigh et al., 1986). By applying a threshold on this HSI value, one can simulate whether a taxon is present or absent. The HSI is an aggregated suitability

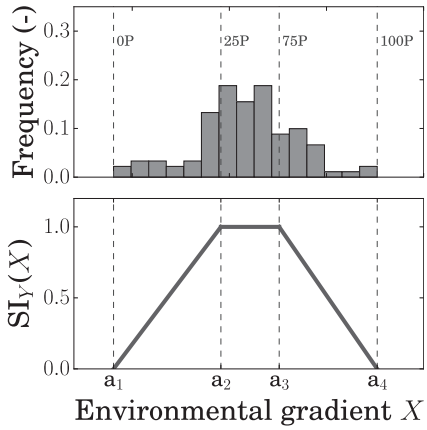

Fig. 2. Estimation of the species response curve for taxon Y along the abiotic gradient X. In the upper panel, the frequency distribution of the measurements of an abiotic variable (e.g. dissolved oxygen) for which the taxon (e.g. Baetidae, family level) is observed is shown. The 0, 25, 75 and 100 percentiles are shown and used to parametrise a_1, a_2, a_3 and a_4 of the species response curve (see lower panel). The SI_Y(X) is the suitability index for abiotic gradient X and taxon Y.

index, quantifying the overall quality of the habitat. Different aggregation functions can be chosen: minimum aggregation (or one-out-all-out principle), averaging, etc. The aggregation or interaction function chosen in this paper is the geometric mean, the m root of m SI values:

$$\mathrm{HSI}_k = \prod_{j=1}^{m} \mathrm{SI}\left(x_k^j\right)^{1/m} \tag{2}$$

With HSI_k being the habitat suitability index of point k. We have chosen this approach because the different constructed species response curves are considered as complementary factors explaining species distribution, rather than redundant factors (highly correlated factors are omitted from the analysis, see Appendix A). This avoids that non-suitability for one abiotic feature is compensated by suitability for another abiotic feature (Langhans et al., 2014). This approach is in contrast with current regression models (for instance Generalized Linear Models (GLMs)), mostly relying on additive models of environmental variables (Termansen et al., 2006).

3.4. Model identification

An SGA is used to identify alternative model formulations for the SDMs. In this study, an SGA is used and implemented so to perform IVS. A main point in the Python implementation of the SGA is that it should allow for extensions to perform parameter estimation and selection of aggregation functions. In the SGA, an alternative structural hypothesis of the conceptual model is searched by using evolutionary operators. The data structure of the genotype is a binary string, encoding either the presence (1) or absence (0) of an input variable. The genotype-phenotype mapper translates the genotype by interpreting a one (zero) as presence (absence) of an input variable in the model. In the objective function, the model is run with the selected variables, and the model output is tested to presence/absence data.

3.4.1. SGA

We used an SGA with three basic operators (selection, crossover and mutation). The tournament selection method of Goldberg and Deb (1991) is applied to select the fittest individuals from a population of chromosomes. In a tournament selection, two random individuals (chromosomes) are chosen to content in a tournament, and the fittest individual (based on an objective function, see section 3.4.2) is selected. The selection rate, defined as the fraction of the population that survives for the next step of mating, is multiplied with the population size (R) to obtain a number of parents. Next, in the crossover operator, the parents are randomly paired to produce offspring with a certain chance, i.e. crossover rate. If mating does not occur, the parents are replaced in the population. This process is repeated until the number of parents plus the number of offspring is equal to the population size (R). The last operator, mutation, is defined as the probability that a random gene is assigned a new value ($0 \rightarrow 1$ or $1 \rightarrow 0$). This process of selection, crossover and mutation is repeated until the best (i.e. near-optimal) solution (i.e. fittest individual) has not changed over a number of iterations or until a maximum of iterations is reached. Additionally, duplicate chromosomes are removed in every generation and replaced with randomly initiated chromosomes. The choice for the values of the hyper parameters for the algorithm and the accompanied performance is discussed in Appendix C.

3.4.2. Objective function

The Akaike information criterion (AIC) is used to test the candidate models on their goodness-of-fit and penalise model complexity. This criterion gives preference to less complex model when two candidate models give a similar goodness-of-fit (Ellison, 2004). Using this criterion decreases the chance of overfitting data and computational efforts needed for parameter estimation (Mouton et al., 2009). The AIC is computed from the least square regression statistic as in Burnham and Anderson (2002):

$$\mathrm{AIC} = n \log\left(\frac{\mathrm{SSE}}{n}\right) + 2m \tag{3}$$

with n, the number of data points, m, the number of variables considered in the model (i.e. model complexity). In order to correct for small number of measurement points (i.e. 120 sampling locations), a correction to the AIC is formulated (AIC_c) (Burnham and Anderson, 2002):

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2m(m+1)}{n-m-1} \tag{4}$$

The sum of squared errors (SSE) is calculated by applying equation (5):

$$\mathrm{SSE} = \sum_{k=1}^{n} \left(\mathrm{P}_k - \mathrm{HSI}_k\right)^2 \tag{5}$$

With k, the index of the data point, P_k, the presence/absence (1/0) of a taxon and HSI_k, the simulated habitat suitability index.

3.5. Model evaluation

Model evaluation is considered as a key aspect in species distribution modelling, since it can give an insight in how well the model performs and identify possible over and under prediction of the developed models (Fielding and Bell, 1997; Manel et al., 2001; McPherson and Jetz, 2007; Mouton et al., 2010). The developed SDMs are evaluated with 7 evaluation measures: the correctly classified instances (CCI), Cohen's Kappa (Kappa), sensitivity (Sn), specificity (Sp), true skill statistic (TSS) and area under the receiver operating characteristic (ROC) curve (ROC-AUC or AUC). The Kappa and TSS are statistics that measures inter-rater agreement for categorical items, normalising the accuracy of a model by the accuracy that might occur by chance alone. The Sn and Sp are the probabilities that the model will correctly classify respectively presence and absence (Fielding and Bell, 1997; Mouton et al., 2010; Gobeyn et al., 2016).

$$\mathrm{CCI} = \frac{\mathrm{TP}+\mathrm{TN}}{n} \tag{6}$$

$$\mathrm{K} = \frac{(\mathrm{TP}+\mathrm{TN})/n - [(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN}) + (\mathrm{FN}+\mathrm{TN})(\mathrm{FP}+\mathrm{TN})]/n^2}{1 - \left[(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN}) + (\mathrm{FN}+\mathrm{TN})(\mathrm{TN}+\mathrm{FP})/n^2\right]} \tag{7}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}} \tag{9}$$

$$\mathrm{TSS} = \mathrm{Sp} + \mathrm{Sp} - 1 \tag{10}$$

with TP, true positive, FP, false positive, FN, false negative, TN, true negative (see confusion table, Table 2). The calculation of CCI, K, Sn,

Table 2
Confusion matrix. TP is true positive, FP, false positive, FN, false negative, TN, true negative.

	Observed	
	Presence	Absence
Predicted		
Presence	TP	FP
Absence	FN	TN

Sp and TSS is based on the confusion matrix and consequently requires a classification of the model output. This is typically done by setting a threshold on the continuous HSI value and classifying the taxon as present or absent. In this paper, the threshold for which the TSS is highest is used as threshold. The AUC is a last measure used in this paper. The ROC curve is a method to visualize the performance of a binary classifier system (see Table 2) as the threshold is varied. (Bradley, 1997; Fawcett, 2006; Fielding and Bell, 1997; Mouton et al., 2010).

Noteworthy is that the measures are used as measure that evaluate the model performance after model development, and not during model identification, as done with the objective function in section 3.4.2. To avoid misleading terminology, the term model evaluation is referred to as the assessment of the final model, whereas performance assessment is referred to as model performance assessment during model identification (Mouton et al., 2010). The objective function described in section 3.4.2 will be used to evaluate and compare the SGA algorithm, whereas the acquired evaluation measures are used to evaluate the developed models.

3.6. Results

In this section, the results of the IVS with SGA for 14 macroinvertebrate taxa are presented. In the first part of this section, the estimated species response curves are shown and shortly discussed (section 3.6.1), followed by the presentation of the results of the IVS with the SGA (section 3.6.2).

3.6.1. Species response curves

Fig. 3 shows the estimates of the species response curves of Baetidae for 12 abiotic variables. The constructed curves are based on empirical data and are subject to data uncertainties. As indicated in section 3.2, different data samples are used for model construction and identification. Different data samples will lead to varying results. This is why the resampling approach is repeated a number of times (see section 3.2). For each of these data samples, the curves are estimated and plotted. This serves as an indication about the uncertainty on the estimated curve.

The curves for Baetidae describe a number of skewed and symmetric responses, defined over either a narrow or large part of the environmental range (Fig. 3). One can observe that a number of curves show a skewed response. For instance, the species response for turbidity, COD, NO_3^--N, NH_4^+-N are skewed towards conditions characterised by no pollution. Only the response curves for DO, temperature, pH and to some extent COD show a symmetric response. In addition, the response is defined over a wide range of conditions for a number of curves. The range for the variable temperature almost completely overlaps with the total measured range. This is not the case for a number of variables characterising pollution, e.g. NH_4^+-N and COD. When inspecting curves for a very sensitive taxa, Gomphidae (Fig. 4, TS = 10), one does not observe a skewed response for pollution-indicating variables, but rather a very narrow response (Fig. 4): chlorophyll, Cl^-, COD, NH_4^+-N, turbidity). Only for temperature, velocity and elevation, the curves are defined over a broad range. This analysis suggests that the taxon requires specific non-polluted conditions to survive.

It is important to note that different data samples lead to different response curves. For Baetidae, the estimate of the response to elevation has a relative high variation over the data samples. This is due to limited measurements in the higher region (only 10% of the measurements are taken above 200 m). This indicates that efforts should be made to measure in the higher parts of this region. Visual inspection of the curves for taxa with low sample prevalence (see Appendix B) indicates that the estimations of the response curves for Corydalidae (prevalence = 9%, Figure B6), Dytiscidae (prevalence = 11%, Figure B7), Gomphidae (prevalence = 14%, Fig. 4), Hydrophilidae (prevalence = 11%, Figure B9) and Hydroptilidae (prevalence = 10%, Figure B10) have large variations. When inspecting the curves for these taxa with low sample prevalence, typically for at least four of the 12 variables a large degree of uncertainty is observed. A low precision of estimated response curves for less sampled taxa is also observed by Karl et al. (2002). Here, however, this is not true for all taxa with a low sample prevalence (e.g. Caenidae, prevalence = 10%), indicating that lesser observed (and possibly rare) taxa do not necessarily have a larger uncertainty in the estimated curves than more observed taxa (generalists). This means that the limited number of presences for one taxon are all observed in the same range of the environmental continuum. Either the taxon prefers these specific conditions or insufficient data are collected over the gradient. In this case study, we suspect that the latter is valid for a number of variables. For example, for NH_4^+-N, a large range of the continuum has not been observed (see Figure A5 in Appendix A). Collecting additional information in these regions of the continuum can give a better insight in the species response and the coupled uncertainties.

3.6.2. IVS with SGA

An SGA (see section 3.4.1) is applied to identify alternate models for the 14 macroinvertebrate taxa. In Appendix C, the methodology for the choice of the hyper parameters is explained. In addition to this, the results of the algorithm performance is tested by repeatedly running the algorithm and comparing the results with a grid search approach. The results of this test are found in Appendix C. From these tests it is concluded that in 99%, the algorithm finds the optimal solution faster than a grid search approach, however, in the remaining 1% a near-optimal solution is found. Similar conclusions are found for a test for the taxon Chironomidae. That is why the hyper parameters (R = 24, crossover rate = 100%, mutation rate = 20%, selection rate = 100%, elitism is used) found for Baetidae are also used for the other taxa. The algorithm is run for a maximum of 50 generations or when the value of the objective function has not changed over 30 generations.

In order to account for the effect of uncertainty in the ecological data, the SGA analysis is repeatedly run with different samples of the data for model construction and identification, similarly to the approach in section 3.6.1. In total, 300 runs are required for each taxon, since this is approximately the required number for convergence of the solution (see Appendix D). For each of these runs, the found (near-)optimal solution is retained, thus resulting in 300 models for each taxon. The structure of the 300 models are analysed by computing the support for variable inclusion, calculated by the dividing the number of runs in which a variable is present in the found (near-)optimal solution by the total number of runs (300).

Fig. 5 shows the mean and standard deviation values of the evaluation criteria for the 14 taxa. The benchmark model, calculated by running the model with all input variables, is also

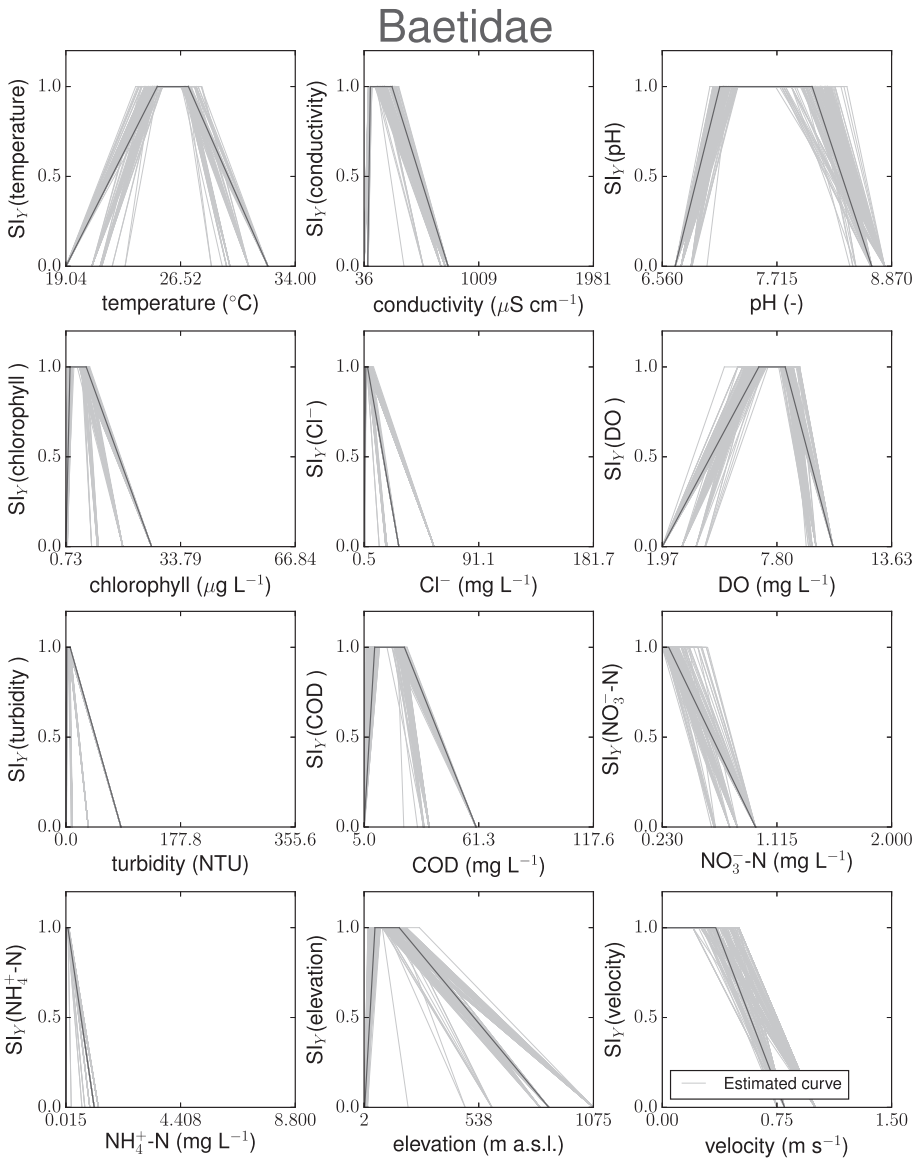

Fig. 3. Species response curves for the taxon Baetidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

evaluated. In general, the IVS with SGA improves the predictive power of the model. The values for Kappa, CCI, TSS and AUC increase - averaged over the 14 taxa - respectively with a value of 0.34, 0.29, 0.27 and 0.19. Following the guidelines of Manel et al. (2001) and Gabriels et al. (2007) a model has a satisfying performance when the Kappa is higher than 0.4 and the CCI and AUC are higher than 0.7. Based on the mean CCI, 11 SDMs have a satisfactory performance, whereas for mean Kappa and mean AUC, only 6 and 5 fulfill the criteria for satisfactory performance. No models have a poor performance (based on Kappa < 0.2) or no discrimination ability (based on AUC < 0.5) (Gabriels et al., 2007; Boets et al., 2013). For one taxon, Libelliludae, the models have approximately on average the same performance as the benchmark model (based on CCI = ± 0.65, Kappa = ± 0.25, AUC = ± 0.7, TSS = ± 0.25). When closely inspecting the results of Libellulidae, one observes that the AIC penalises complex models leading to a higher performance. This suggest that in this case the implemented AIC might over-penalise more complex models. In contrast, it is not uncommon that species presence/absence patterns are characterised by a limited number of factors (Boets et al., 2013; Lock et al., 2014).

The values in Fig. 5 are reported for a HSI threshold for which the TSS reaches a maximum. Since each criteria has its own mathematical formulation, the threshold for which these criteria reach a maximum will be different. In addition, the Kappa and CCI depend on the prevalence of the used data, whereas TSS and AUC are assumed to be independent (Allouche et al., 2006; Mouton et al., 2010). In order to limit the bias due to prevalence in the analysis, the threshold which maximizes the TSS is chosen. Since the prevalence for all 300 models per taxa is kept constant in the data resampling, the comparison of the evaluation criteria for these models is justified. However, one has to be careful when comparing SDMs for different taxa with Kappa and CCI, since the reported criteria vary as a function of the taxa prevalence.

In Figs. 6–8, the support for variable inclusion is found for the 14 taxa. The Shannon entropy (Shannon, 1948) on the support is calculated and used as a measure of uncertainty. In this measure, the uncertainty on the support is minimal (i.e. zero) when the SGA selects a variable in all or none of the runs. The uncertainty is

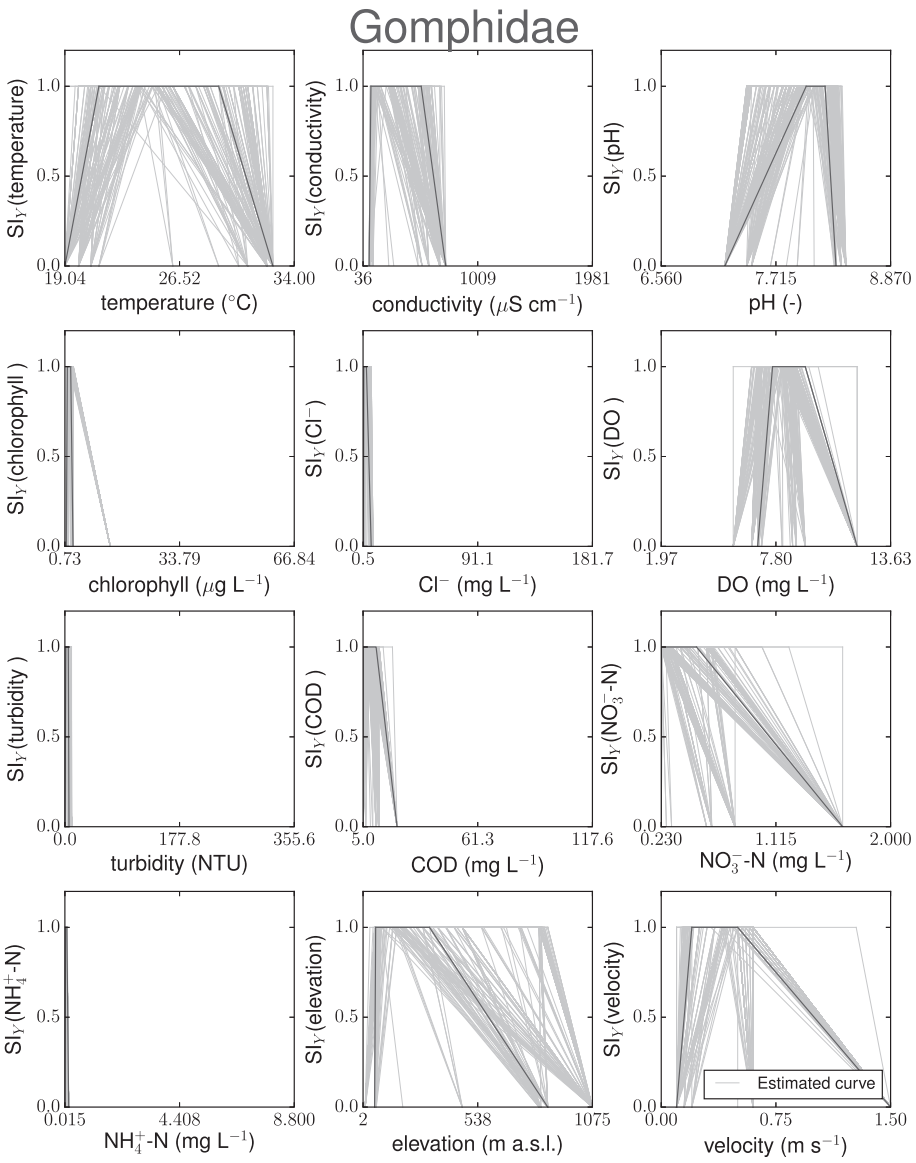

Fig. 4. Species response curves for the taxon Gomphidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

maximum (i.e. one) when 150 SGA runs selected a variable, and the other 150 runs did not. In general, a limited number of variables is included in the models, and if they are included the uncertainty on their inclusion is generally high. For Acari, Coenagrionidae, Libellulidae and Thiaridae, respectively one (elevation), two (elevation and chlorophyll), one (elevation) and three (pH, elevation and conductivity) variables have (has) a high support, whereas for the other taxa, the identification of a limited set of explanatory variables is characterised by uncertainty (e.g. Corydalidae, Gomphidae, Hydroptilidae and Tubificidae). For some taxa, it is straightforward to identify which variables are not explaining species presence/absence patterns. For instance, for Baetidae, it is observed with relative certainty that velocity, pH, NH_4^+-N, conductivity, Cl^- and COD do not explain the presence/absence patterns in the data.

4. Discussion

In this paper, we suggest the use of EAs to identify alternative models for an SDM. Therefore, an SGA is used as a tool to perform IVS for models of 14 macroinvertebrate taxa. In section 4.1, the relevance of the with SGA identified alternative models to investigate river pollution will be discussed. In section 4.2, the performance and use of SGA for IVS is discussed, whereas in section 4.3, a general discussion about the perspective of EAs in model identification for SDMs is presented. Finally, we briefly discuss the implemented code which is used for this study (section 4.4).

4.1. Case study

A series of macroinvertebrate SDMs are developed by using species response curves. These curves can provide a first insight in which variables possibly explain the biological response of the taxa. With the SGA a set of the explanatory variables for species presence/absence patterns are identified. These two elements provide information which can aid ecologist to obtain an increased system insight and inform decision managers with possible effects of management actions. Although the developed models can be used for this purpose, it is important to note that the models can be

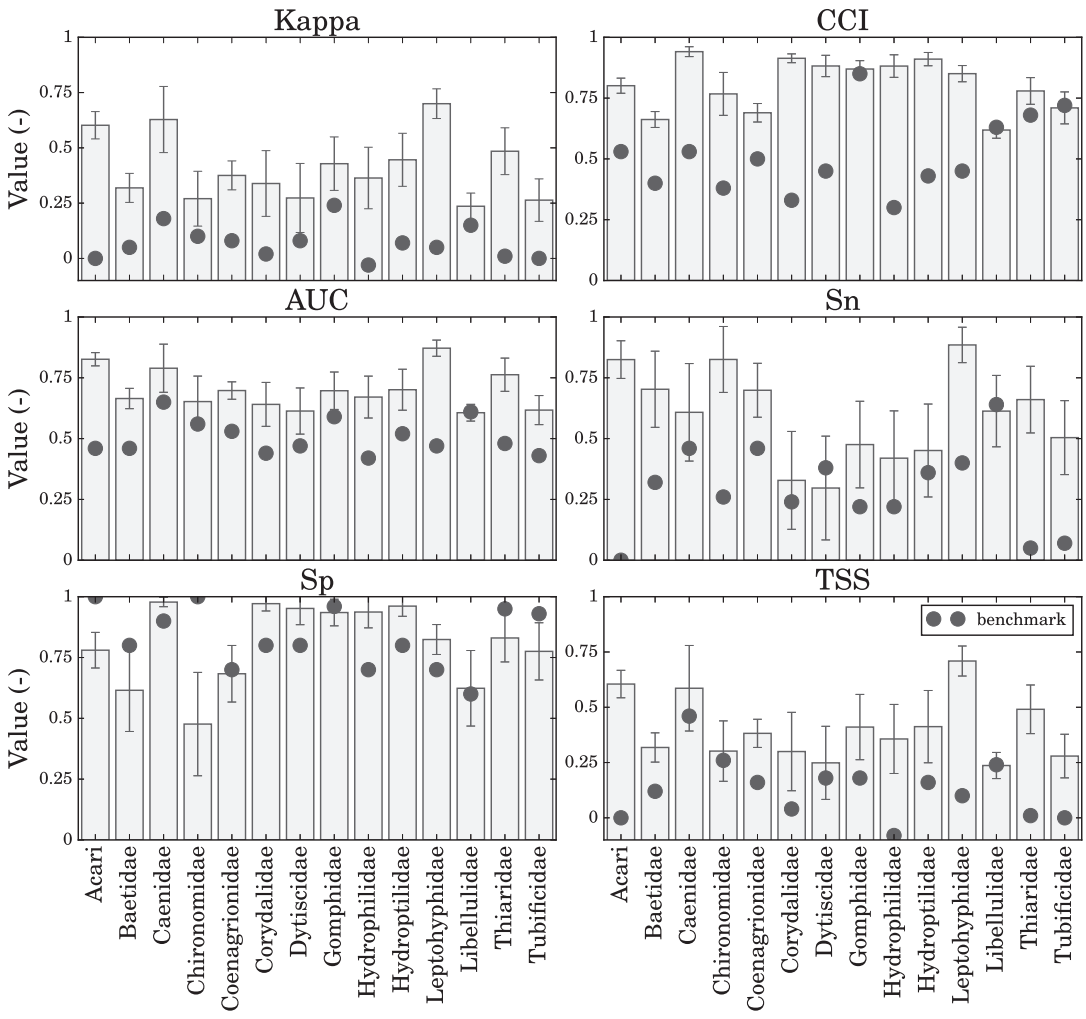

Fig. 5. Evaluation with Kappa, CCI, AUC, Sn, Sp and TSS of the models identified with the SGA. In total 300 models per taxon are summarized (bar graphs). The benchmark model (model with all variables) is also evaluated and given by the dark grey dots. The standard deviation on the mean is given as a measure of uncertainty. The measures are reported for the threshold which gives the maximum performance for TSS.

further optimised in terms of accuracy and precision. In addition, only 14 taxa are considered in this study, whereas for a complete overview, an estimate should be made for more taxa.

In order of importance, elevation, velocity, conductivity, chlorophyll, pH, NH_4^+-N and DO are identified as explanatory variables. The rather low importance of nutrient variables is in contrast with the hypothesis set forward by Ambarita et al. (2016), especially since the estimated effects for some pollution sensitive taxa (i.e. Baetidae, Gomphidae, Leptohyphidae) are not explained by eutrofication indicating variables. The importance of elevation could be explained by the pooled character of the variable; it can be assumed that the elevation pools the effect of pressures on the system (upstream: less intense agriculture, downstream: urban and intense agricultural activity). A complete overview of the system can be given by running the analysis for all taxa, given enough information is available. Yet, this was out of scope of the current study. In the following, we provide and discuss suggestions for future studies to increase (decrease) reliability (uncertainty) of the SDMs.

Data quality and quantity are of major relevance for model identification. It is expected that an increase of the number of measurements will increase model precision (see Hernández et al. (2006)). However, Engler et al. (2004) does indicate that data quality might be of bigger importance, especially for the development of models for rare taxa. An improved field data collection strategy can increase quality of the data. One could investigate how measurements can be taken so they better reflect the abiotic conditions for the studied taxa, based on spatial and temporal dynamics of the variable and characteristics of the taxon at interest. In addition, one should aim to sample the study area along a pollution gradient. This implies that an equal share of polluted and non-polluted sites should be sampled. This will decrease the risk of introducing sample bias in the SDMs (Fithian et al., 2014). Data resampling can partly solve this issue (Guisan and Zimmermann, 2000), given that enough relevant information remains after resampling.

Uncertainties in the model structure identification can also be reduced by refining model construction and identification process with expert, process and ecological knowledge (for examples, see case studies of Everaert et al. (2013); Fu and Guillaume (2014) and Bennetsen et al. (2016)). With respect to the latter, information from an ecological information database (for example, the freshwaterecology.info database of Schmidt-Kloiber and Hering (2015)) can aid to set boundaries for the model identification.

4.2. SGA as a tool for IVS

The SGA used and implemented for this case study is able to identify (near-) optimal solutions in a shorter time frame than a grid

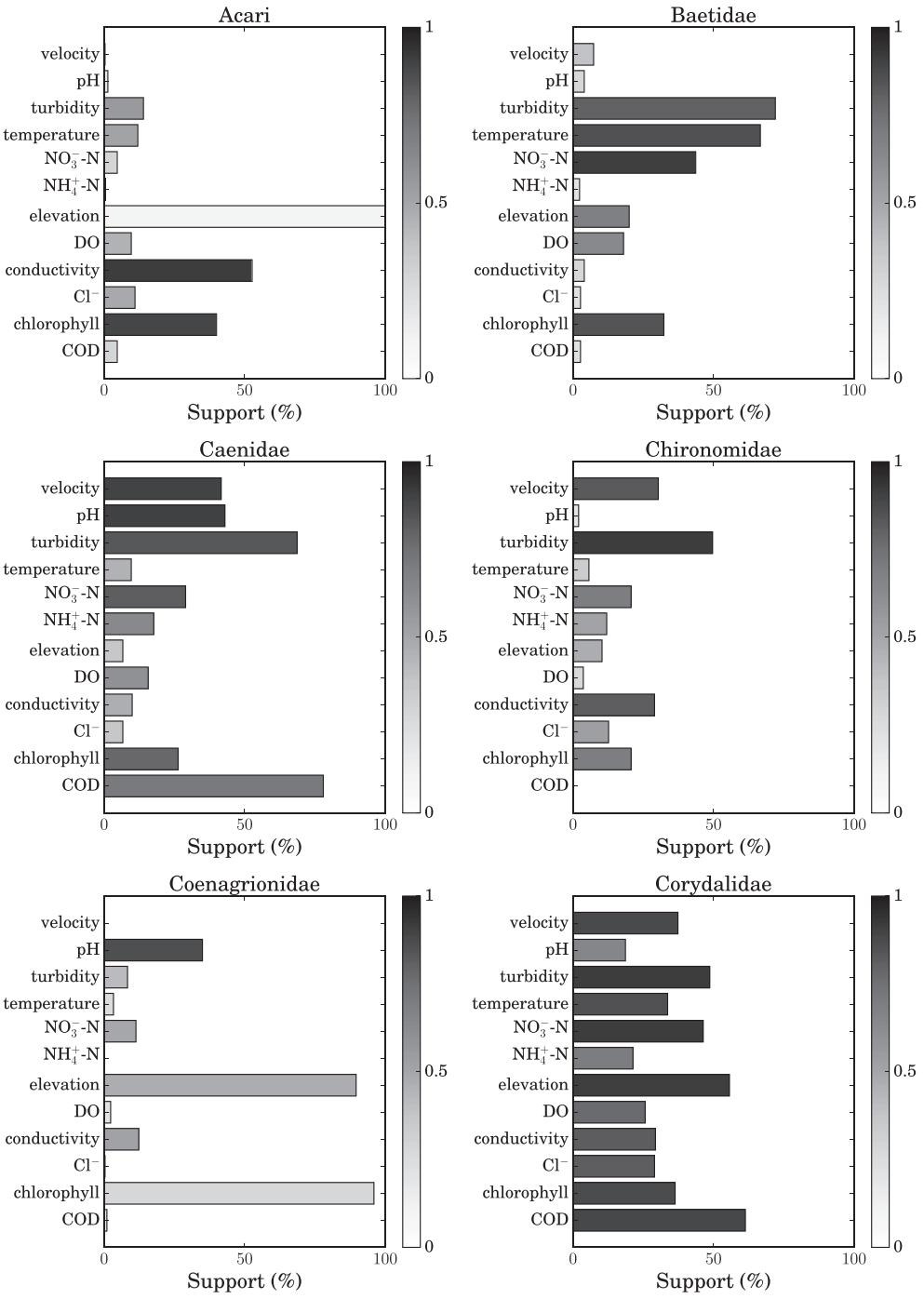

Fig. 6. Support for the variable inclusion for 300 repeated analysis for Acari, Baetidae, Caenidae, Chironomidae, Coenagrionidae, Corydalidae. The uncertainty (greyscale) is estimated by calculting the Shannon entropy.

search approach. The computational time can be reduced, however, this is at expense of the guarantee that the optimal solution will be found. This presents a trade-off between available computational time and the necessity for the optimal solution. With this in mind, it is important to note that the search space in this study is rather small (2^{12}-1 = 4095 options). For modelling of freshwater taxa, the problem size typically ranges between 10 and 20. D'heygere et al. (2003) use seven physico-chemical and five hydromorphological variables to analyse the habitat preferences of eight generalist macroinvertebrate taxa in Flanders. In this approach, an SGA is used as a wrapper to perform IVS for decision trees to increase model performance. Similarly, Boets et al. (2013) use nine physico-chemical and one hydromorphological variables as input for by an SGA optimised decision tree to analyse the habitat and the invasion risk of *Dikerogammarus villosus*. In another example, Domínguez-Domínguez et al. (2006) use 14 variables, describing the topography and climate, as input for the GARP methodology in order to test the conservation status of goodeines, a group of viviparous freshwater fishes in central Mexico. Further, Bennetsen et al. (2016) use 34 variables as input for a conceptual SDM. The authors

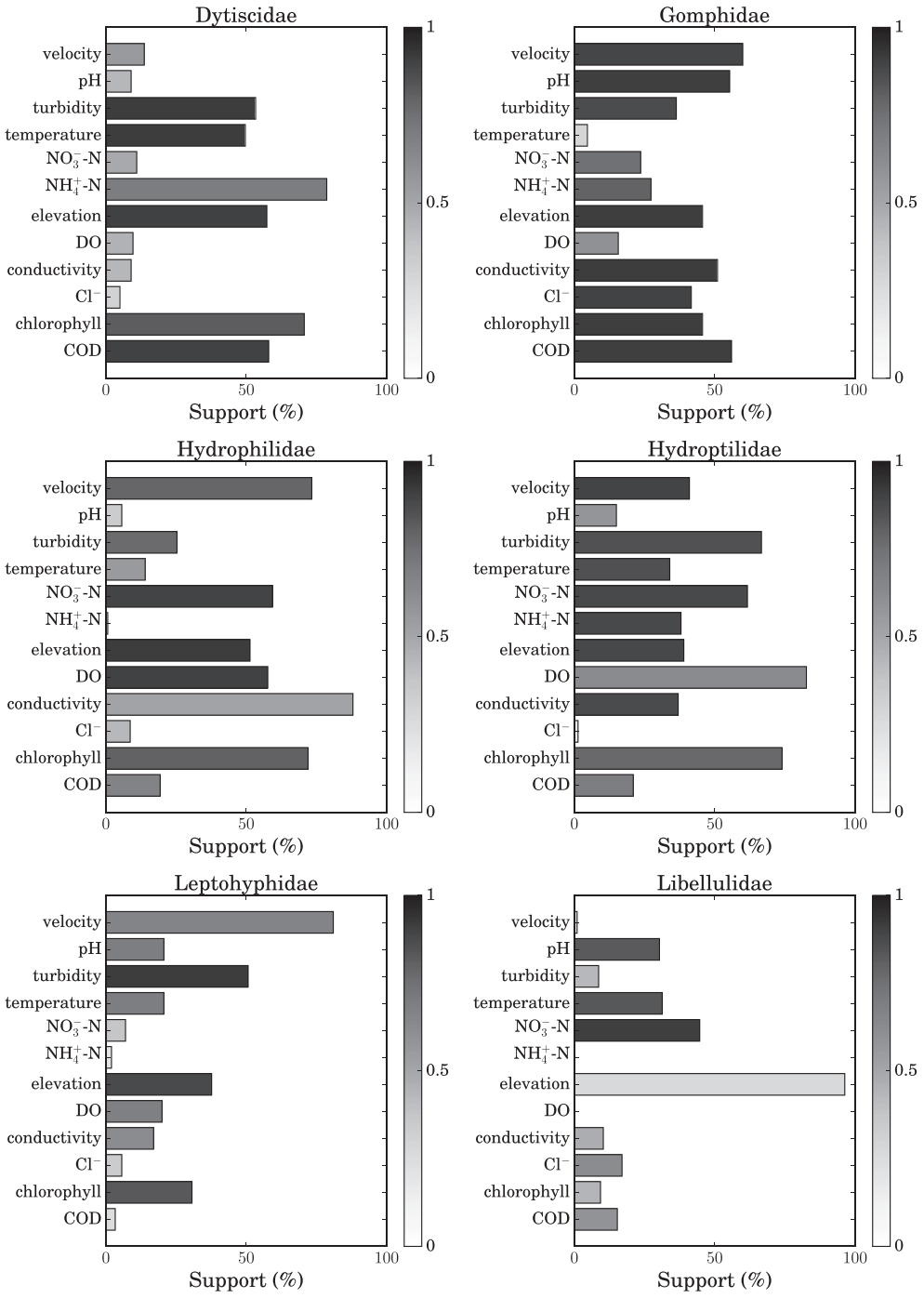

Fig. 7. Support for the variable inclusion for 300 repeated analysis for Dytiscidae, Gomphidae Hydrophilidae, Hydroptilidae, Leptohyphidae, Libellulidae. The uncertainty (greyscale) is estimated by calculting the Shannon entropy.

conclude that the model complexity is a potential flaw of the model and tools reducing this complexity can aid to improve model performance. The SGA implemented for this case study can be used as a tool, however, it will be important to verify its performance. With the results at hand, it is suspected that for larger problems (>15), the SGA will not be able to identify the global optimal solution within a reasonable time frame, but rather a near-optimal solution. In this case, adaptation to the SGA can be made, where a local search algorithm, like Hill Climbing (HC), can aid to improve the exploitative capacity (local searching) of the algorithm.

A number of IVS strategies are available in literature (May et al., 2011). With respect to species distribution modelling, back- and forward variable selection approaches or a combination of both are generally used (for example, see Mouton et al. (2009)). In this process, an alternative model is tested to data by iterative excluding (or including) a variable from the model (Zuur et al., 2010). These approaches are considered as greedy, because they make local optimal decisions with the assumption that the global solution will be found. Although the forward selection approach is computational efficient, it only searches for a small subset, and

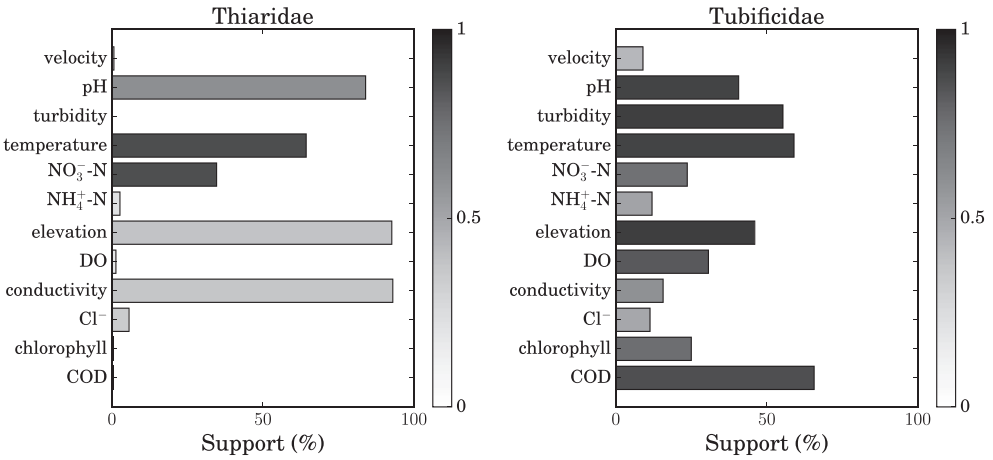

Fig. 8. Support for the variable inclusion for 300 repeated analysis for Thiaridae and Tubificidae. The uncertainty (greyscale) is estimated by calculting the Shannon entropy.

may encounter local optima rather than global. In addition, this procedure may ignore informative combinations of input variables which are only marginal relevant individually. A disadvantage of backward selection is the difficulty to determine the relative importance of an input variable compared to forward selection. Moreover, the backward search might be biased towards larger models (May et al., 2011). In this perspective, heuristic search methods, as SGA, are attractive options to efficiently search for the global (or near-global) optimum (May et al., 2011; Galelli et al., 2014; Mount et al., 2016). As stated above, there is still no guarantee that the global optimal solution is found and important, these types of algorithms often require a certain degree of experience, making the use of the tool for IVS not a trivial task. With this, the number of hyper parameters that need to be set is often considered as a disadvantage of SGAs and EAs in general (Reed et al., 2000). There are guidelines that support the selection of the parameters with a number of hierarchical rules (Gibbs et al., 2008), however, the proper choice depends on the problem at hand (Gibbs et al., 2015).

This is potentially the reason why only seven studies have used SGAs to perform IVS for SDMs (based on a scan of Web Of Science, 22/11/2016). All of these studies apply SGAs as a wrapper to perform IVS for data-mining approaches in the context of freshwater management. Three of these studies (D'heygere et al., 2006; Boets et al., 2013; Sadeghia et al., 2013) use a limited iterative approach to find near-optimal values for the hyper parameters, whereas the remainder of the studies (Ambelu et al., 2010; Vayghan et al., 2016; Muñoz-Mas et al., 2016; Zarkami et al., 2014) do not specify their approach for hyper parameter setting. In addition, the convergence criterion is similar to the one selected in this paper, a maximum of 20–40 generations, for a comparable problem size. In none of the studies, the performance of the SGA is evaluated. Together with this study, it shows that SGAs, and EAs in general, have an added value as wrapper to support model identification.

4.3. Perspectives for EAs as a SDM identification tool

In this paper, a conceptual approach is used as the basis for the SDMs. In this perspective, this approach presents an alternative to SDMs mostly determined by data-driven approaches, i.e. machine learning and statistical approaches. However, the conceptual SDMs require a degree of fitting so to test the model to empirical data. Here, a model identification tool based on EAs is proposed.

Model identification is an essential component in model development of (conceptual) SDMs. This process of model identification concerns IVS, parameter estimation and choice of aggregation function. As shown in this paper, SGAs can aid in this process of IVS, but possibly also in parameter estimation and selection of aggregation function. The potential of SGAs, and in general of EAs, in model identification is in the way the problem can be programmed in the genotype of the chromosomes (Eiben and Smith, 2015). The data structure of a genotype is often a binary string, which can be translated by a self-defined genotype-phenotype mapper, mapping binary strings to parameter values, included variables and aggregation functions. In this way, the EAs can be used as embedded method, rather than as a wrapper for data-driven methods (e.g. DT or Artificial Neural Networks (ANN), (D'heygere et al., 2006)).

The potential of the EA approach for SDMs will have to be validated by applying the approach to other case studies. However, its potential should also be tested on a more fundamental level, which researches the fundamental characteristics of the problem at hand (Maier et al., 2014). These characteristics are presented in the fitness landscape, which defines the search space of the optimisation problem. In case of EAs, this fitness landscape will be shaped by what (parameter estimation, IVS or both) and the way the problem (binary, real-valued, grey) is coded. Information that can summarize the characteristics of the fitness landscape could be very valuable to improve and assess the presented approach. Even more, fundamental insights in the search behavior of SGAs, EAs and metaheuristic algorithms in general (e.g. HC, particle swarm optimisation (Van Broekhoven et al., 2006; Mouton et al., 2009, 2011)) could also help improve and offer a perspective on the current approach.

Advances in fundamental knowledge and experimental work with EAs could offer the chance for researchers to test hypothetic ecological concepts related to species dispersal and biotic interactions in models. The inclusion of these concepts can potentially improve the conceptual ground of these SDMs and in addition improve the predictive capacity of the developed SDMs (Boulangeat et al., 2012; Kissling et al., 2012; Wisz et al., 2013; Vezza et al., 2015).

4.4. Available software

The software that we offer with this study can be used as a tool to identify alternative model hypothesis for a conceptual SDM. The code is written so batch run and consequently repeated model

development (e.g. with different data samples) is promoted. We believe this approach offers the opportunity to get better insights in model uncertainty, which is very relevant to uncertain ecological data. For the implementation of the code, the Anaconda Python distribution package (Continuum Analytics, https://www.continuum.io/why-anaconda) is used. All implemented scripts are freely available online, so users can test, apply and change the code.

5. Conclusions

Model identification, defined as the process of parameter estimation, selection of aggregation functions and input variable selection (IVS), is an essential component in species distribution modelling to obtain robust and reliable tools for freshwater decision management. In this study, it is shown that a simple genetic algorithms (SGA), a type of evolutionary algorithm (EA), is suitable tool to perform IVS for SDMs based on habitat suitability and niche and filter theory. Although the coupled uncertainties are relatively high, the process of IVS leads to increased performance of the SDMs, and moreover to a refined insight in the species response to environmental conditions. With this, a tool is developed which is useful to test multiple SDM hypotheses, facilitating the search for factors and interactions shaping the species assemblage.

Acknowledgements

Sacha Gobeyn is supported by a Bijzonder Onderzoeksfonds project related to the Ecuador Biodiversity Network of the Vlaamse Interuniversitaire Raad-Universitaire Ontwikkelingssamenwerking (VLIR-UOS). This research was performed in the context of the VLIR Ecuador Biodiversity Network project. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government - department EW. The authors would also like to thank three anonymous reviewers for their interesting insights and constructive comments on the manuscript. In addition, the authors would like to thank J. Van Butsel and M.A.E. Forio for their valuable comments and critical insight in the early versions of this manuscript. The authors would also like to thank E. Haspeslagh for improving the readability of the text.

Appendix A. Data processing & exploration

The variables considered in this study are listed in Table A1. The values below the detection limit were set to the detection limit. In order to identify collinearity between variables, the spearman rank correlation (r_{spear}) between the variables is calculated (see Appendix A). A threshold value of 0.70 ($> |r_{spear}|$) is set to inspect correlation between variables (Dormann et al., 2013). In this case study, DO_{sat} (correlated to DO) and TDS (correlated to conductivity) are omitted. In addition, box plots, histograms and dot plots are analysed (see Figures A1 until A6) to detect outliers. Four field samples are closely inspected, for which three are assessed as valid, whereas for one field sample, the value of COD is omitted from the analysis.

Table A1
Summary statistic of the variables considered for the modelling exercise. The included variables in the modelling exercise are indicated in the last column.

Variable	Unit	Minimum	Median	Mean	Maximum	Number of records[a]	Included
COD	mg L^{-1}	5.00	13.25	17.02	117.60	100	x
chlorophyll	mg L^{-1}	0.73	3.11	5.59	66.84	120	x
Cl^-	mg L^{-1}	0.53	2.50	7.28	182	120	x
conductivity	μS cm^{-1}	36.50	123.34	199.94	1981	120	x
DO	mg L^{-1}	1.97	7.76	7.50	13.63	120	x
DO_{sat}	mg L^{-1}	23.63	93.20	92.24	178.99	120	
elevation	m a.s.l.	2.00	82.00	135.05	1075	120	x
NH_4^+-N	mg L^{-1}	0.02	0.06	0.20	8.80	118	x
NO_3^+-N	mg L^{-1}	0.23	0.23	0.37	2.00	118	x
TDS	g L^{-1}	0.05	0.08	0.13	1.27	120	
temperature	°C	19.04	26.04	25.98	34.00	120	x
turbidity	NTU	0.00	3.39	9.82	355.63	120	x
pH	–	6.56	7.59	7.66	8.87	120	x
velocity	m s^{-1}	0.00	0.15	0.23	1.50	120	x

[a] The number of records include the records for which the value was below or above the detection limit.

Table A2
Spearman rank matrix.

	temperature	conductivity	TDS	pH	chlorophyll	Cl^-	DO	DO_{sat}	turbidity	COD	NO_3^--N	NH_4^--N	velocity	elevation
temperature	1	−0.08	−0.12	−0.3	0.56	0.39	−0.19	0.1	0.4	0.37	−0.16	0	−0.68	−0.62
conductivity	−0.08	1	0.97	0.35	−0.24	0.35	−0.07	−0.06	−0.05	0.19	0.33	0.13	0.24	−0.24
TDS	−0.12	0.97	1	0.31	−0.24	0.36	−0.09	−0.09	−0.05	0.23	0.28	0.14	0.26	−0.25
pH	−0.3	0.35	0.31	1	−0.38	0.06	0.77	0.68	−0.23	−0.26	0.27	−0.1	0.6	0.4
chlorophyll	0.56	−0.24	−0.24	−0.38	1	0.27	−0.19	−0.07	0.62	0.29	−0.49	0.12	−0.69	−0.34
Cl^-	0.39	0.35	0.36	0.06	0.27	1	−0.03	0.13	0.24	0.17	−0.15	0.16	−0.18	−0.42
DO	−0.19	−0.07	−0.09	0.77	−0.19	−0.03	1	0.93	−0.26	−0.2	0.23	−0.22	0.37	0.43
DO_{sat}	0.1	−0.06	−0.09	0.68	−0.07	0.13	0.93	1	−0.12	−0.11	0.22	−0.26	0.18	0.22
turbidity	0.4	−0.05	−0.05	−0.23	0.62	0.24	−0.26	−0.12	1	0.16	−0.31	0.22	−0.35	−0.34
COD	0.37	0.19	0.23	−0.26	0.29	0.17	−0.2	−0.11	0.16	1	−0.02	0.18	−0.3	−0.4
NO_3^--N	−0.16	0.33	0.28	0.27	−0.49	−0.15	0.23	0.22	−0.31	−0.02	1	−0.37	0.19	0.14
NH_4^--N	0	0.13	0.14	−0.1	0.12	0.16	−0.22	−0.26	0.22	0.18	−0.37	1	−0.07	−0.25
velocity	−0.68	0.24	0.26	0.6	−0.69	−0.18	0.37	0.18	−0.35	−0.3	0.19	−0.07	1	0.53
elevation	−0.62	−0.24	−0.25	0.4	−0.34	−0.42	0.43	0.22	−0.34	−0.4	0.14	−0.25	0.53	1

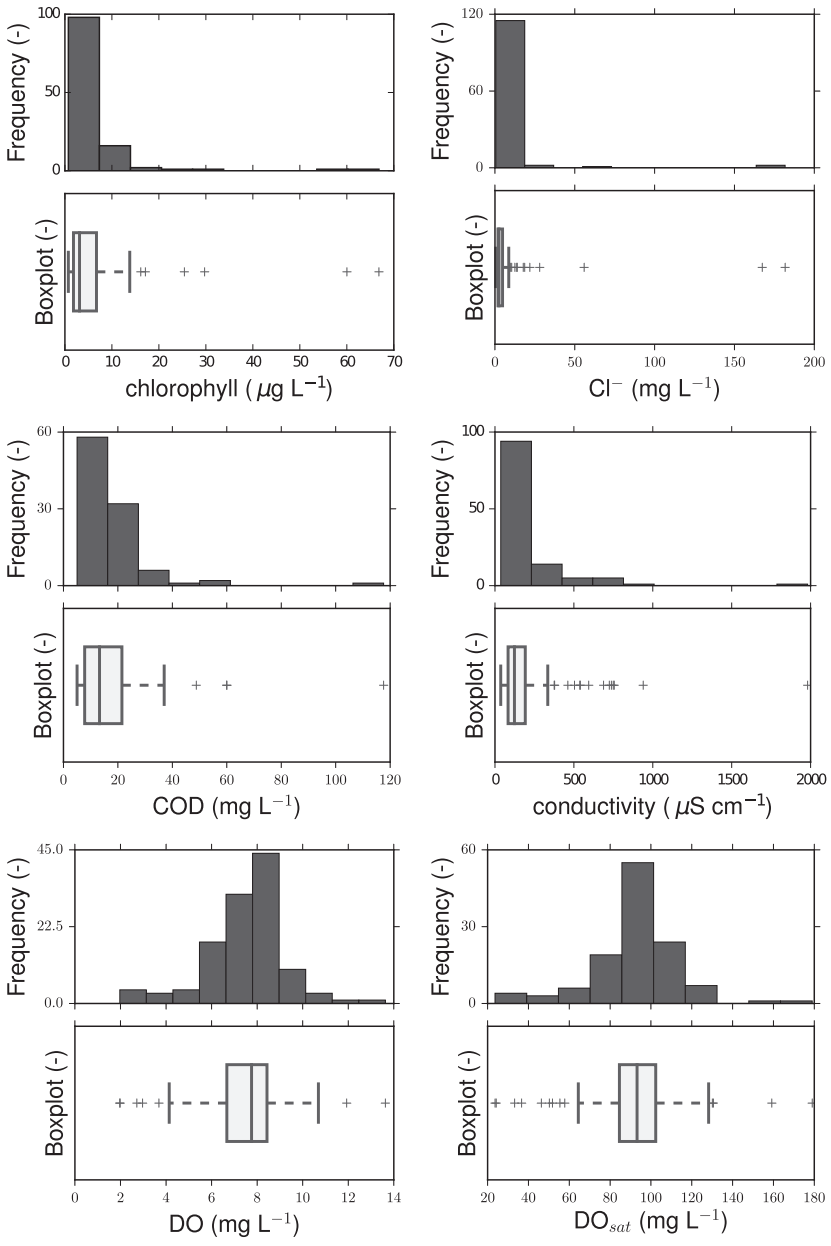

Fig. A1. Boxplots and histograms for the variables chlorophyll, Cl., COD, conductivity, DO and DO_{sat}.

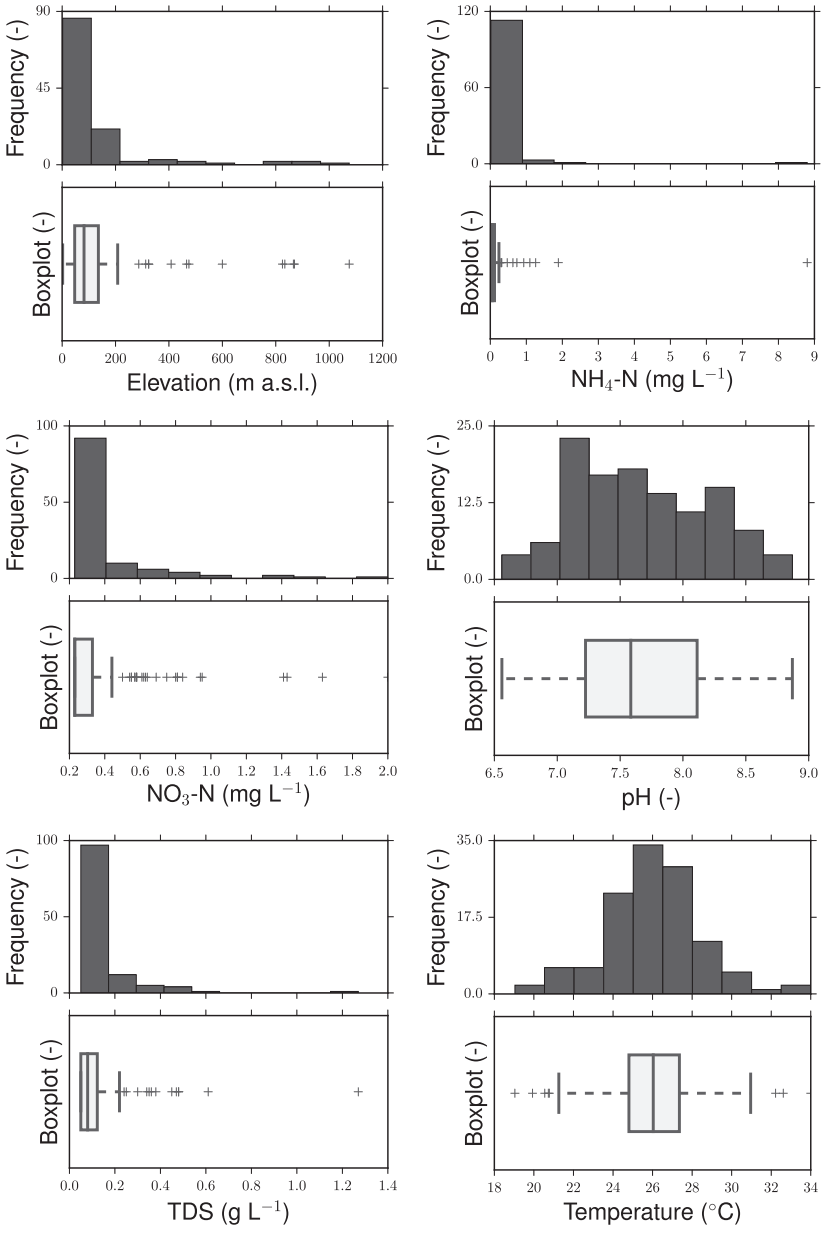

Fig. A2. Boxplots and histograms for the variables elevation, NH_4^--N, NO_3^-, pH, TDS and temperature.

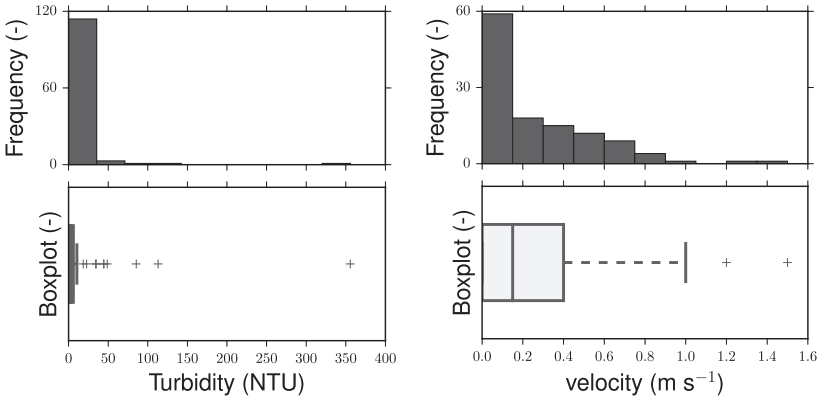

Fig. A3. Boxplots and histograms for the variables Turbidity and velocity.

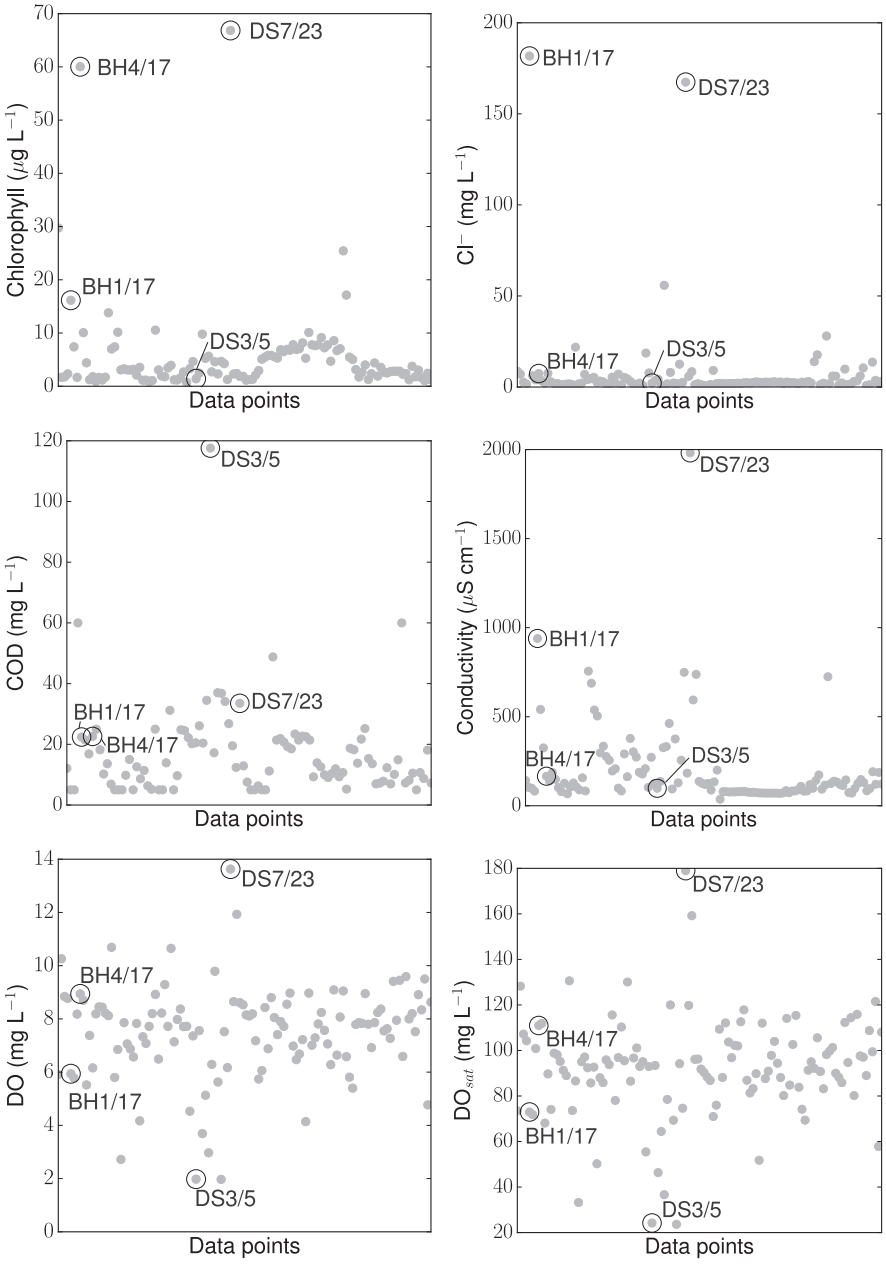

Fig. A4. Dotplots for the variables chlorophyll, Cl^-, COD, conductivity, DO and DO_{sat}. The points which were inspected are indicated in grey.

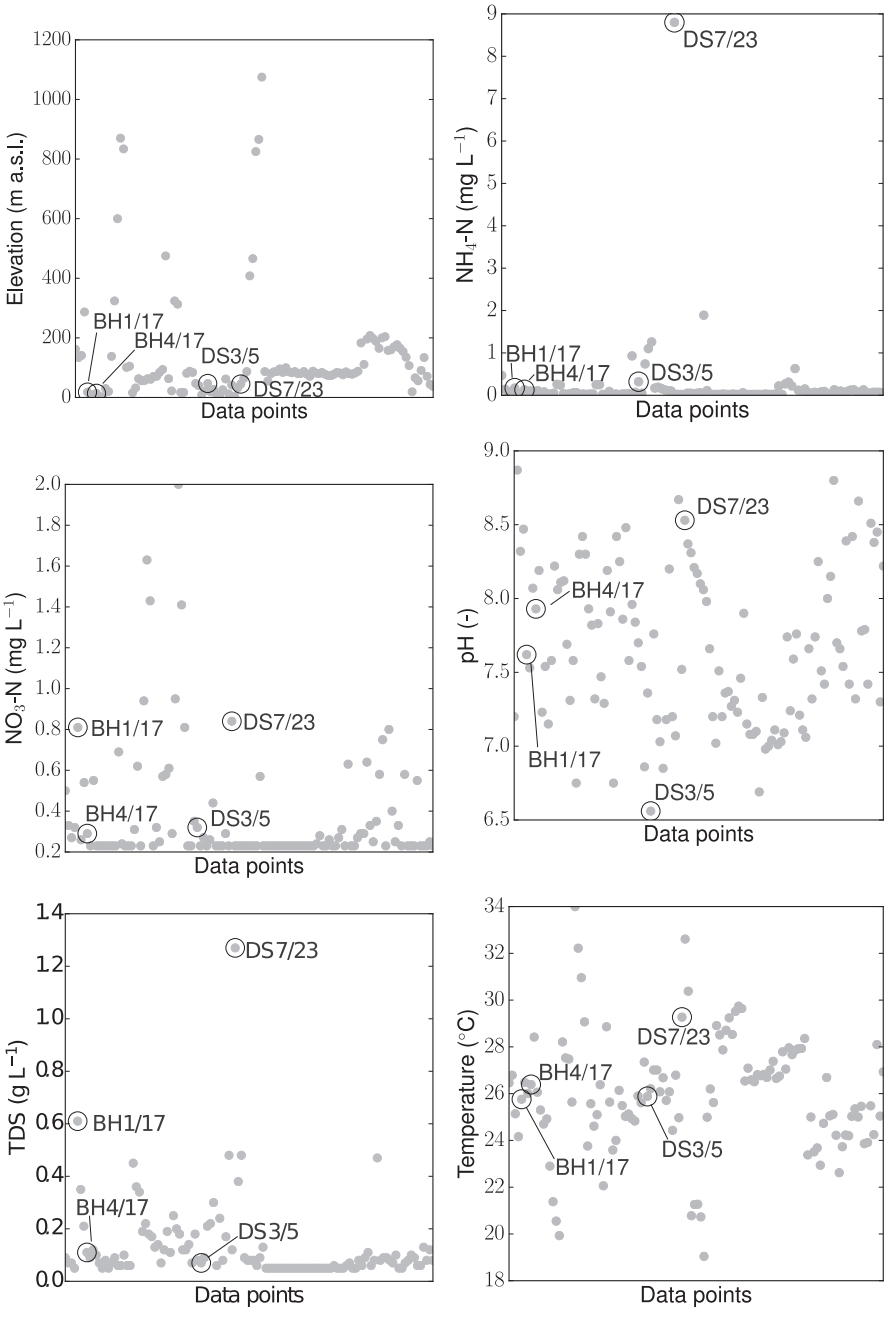

Fig. A5. Dotplots for the variables elevation, NH_4^--N, NO_3^-, pH, TDS and temperature. The points which were inspected are indicated with a circle.

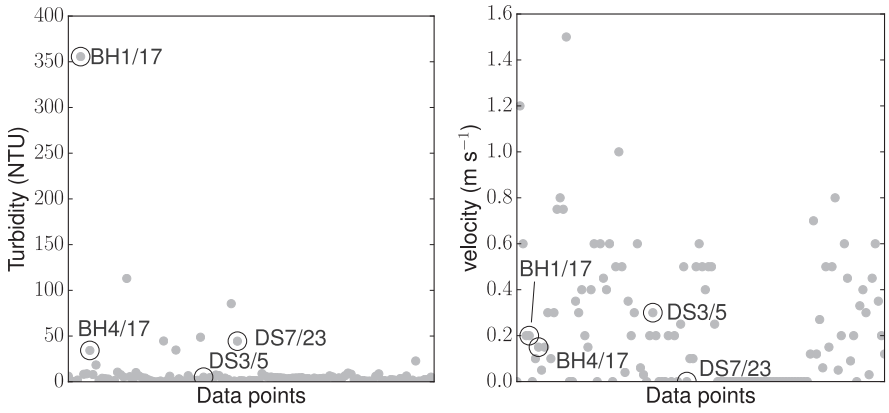

Fig. A6. Dotplots for the variables turbidity and velocity. The points which were inspected are indicated with a circle.

Appendix B. Species response curves

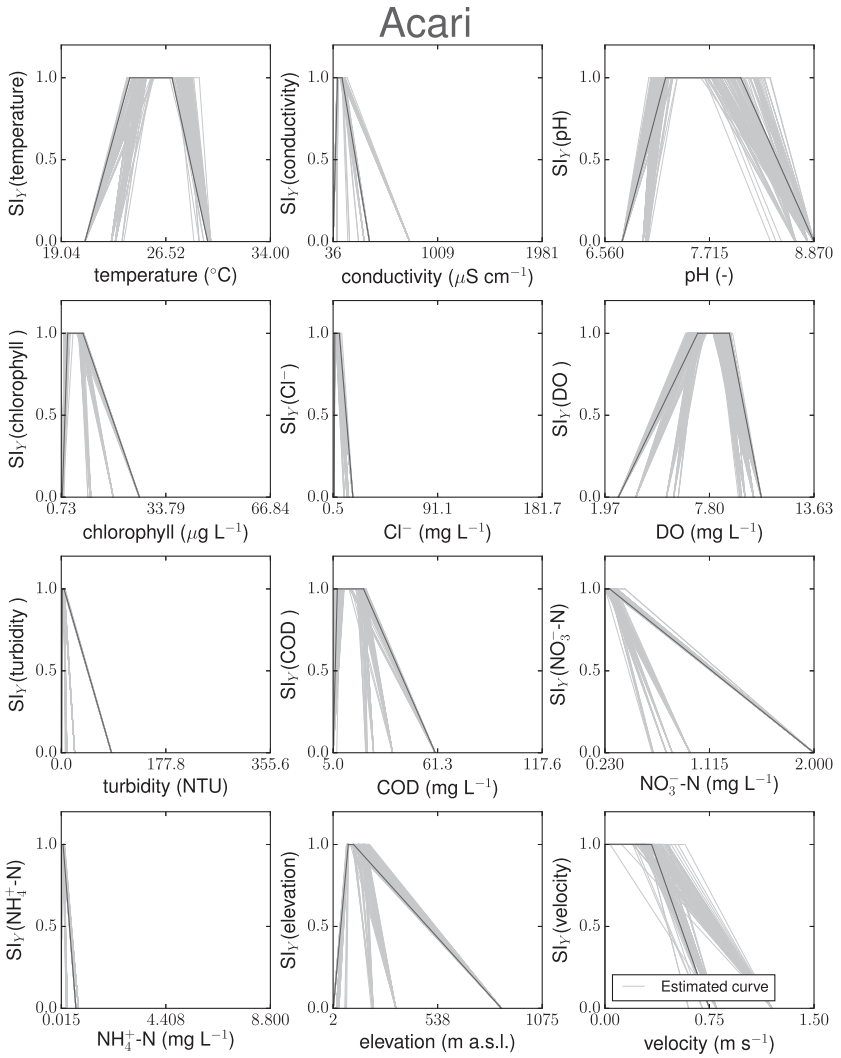

Fig. B1. Species response curves for the taxon Acari for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

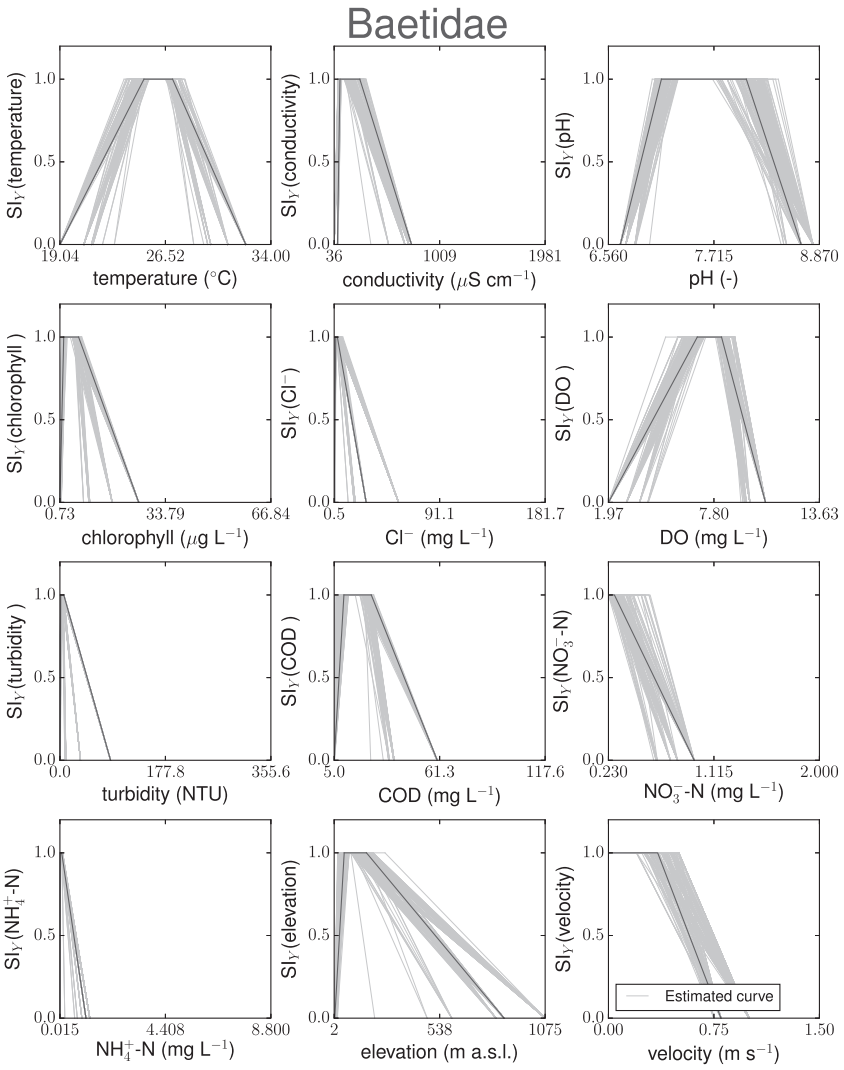

Fig. B2. Species response curves for the taxon Baetidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

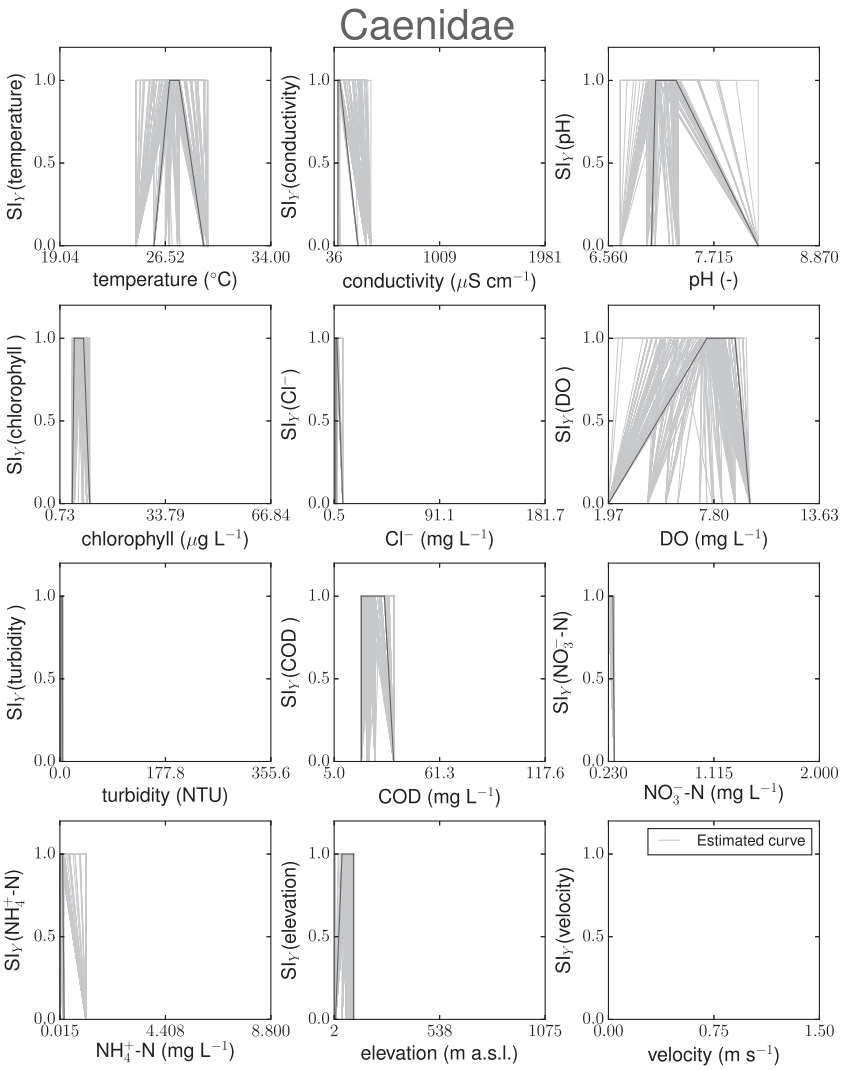

Fig. B3. Species response curves for the taxon Caenidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

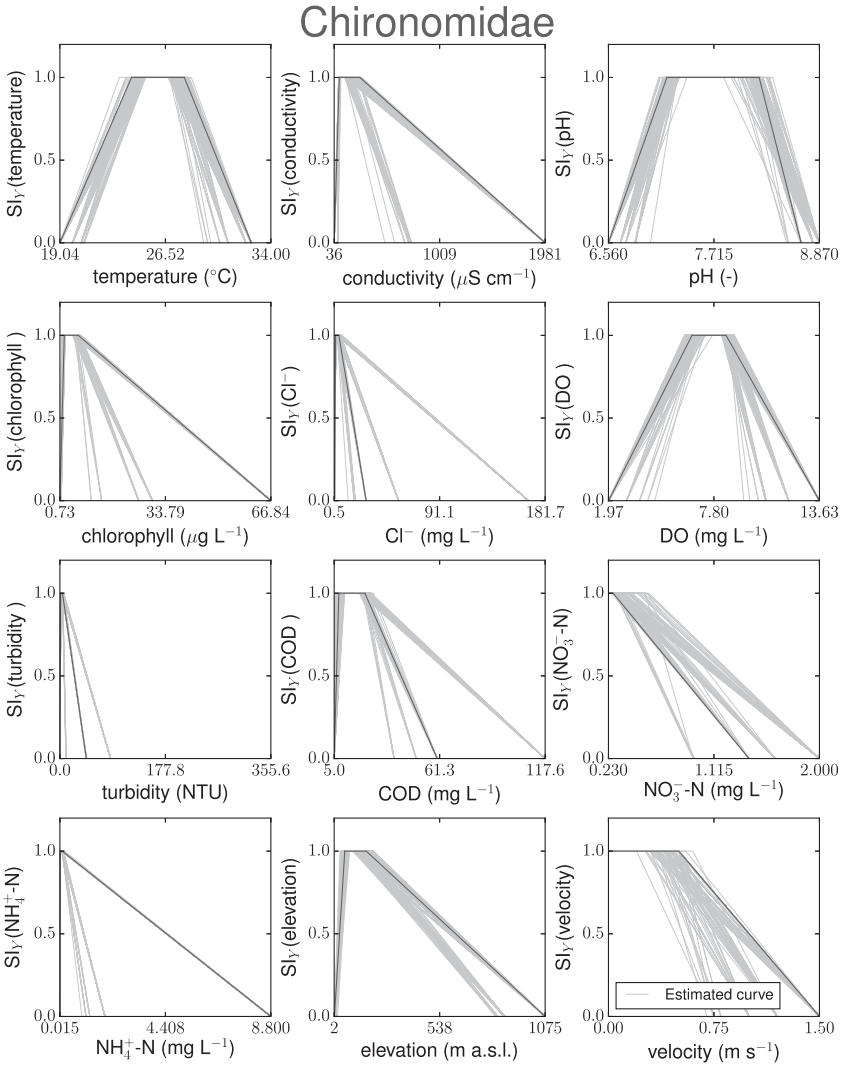

Fig. B4. Species response curves for the taxon Chironomidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

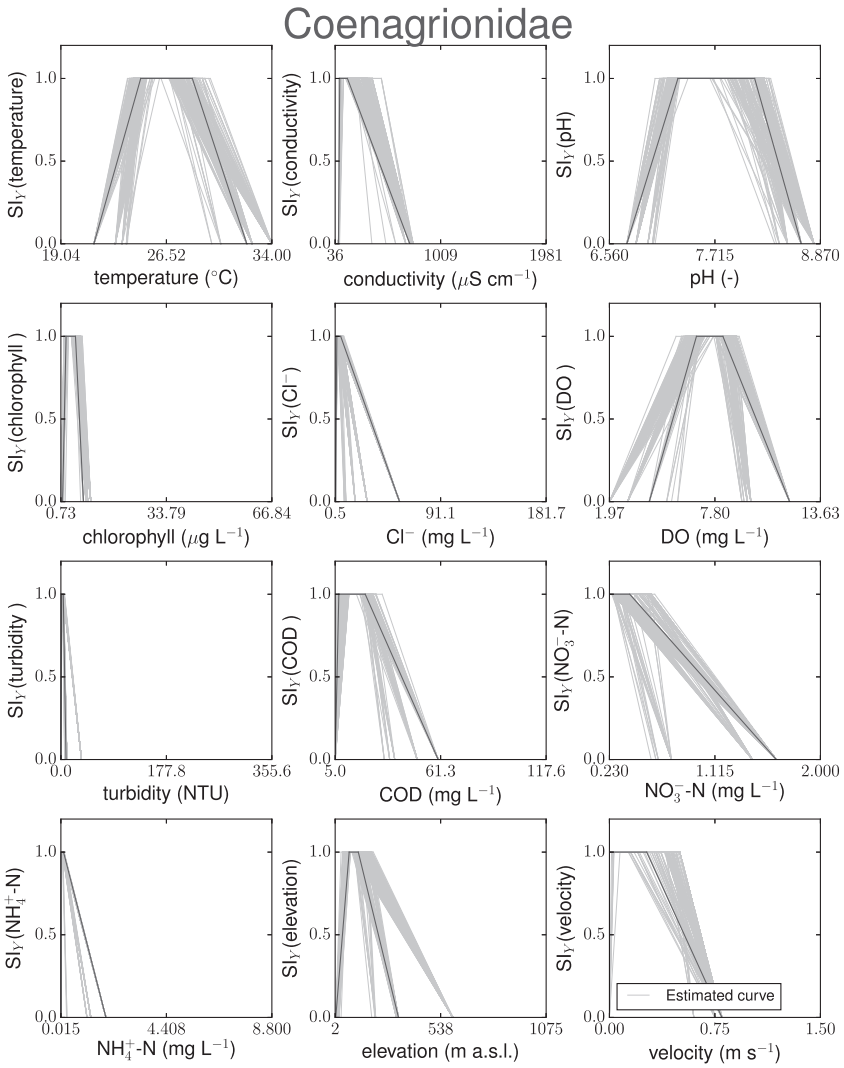

Fig. B5. Species response curves for the taxon Coenagrionidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

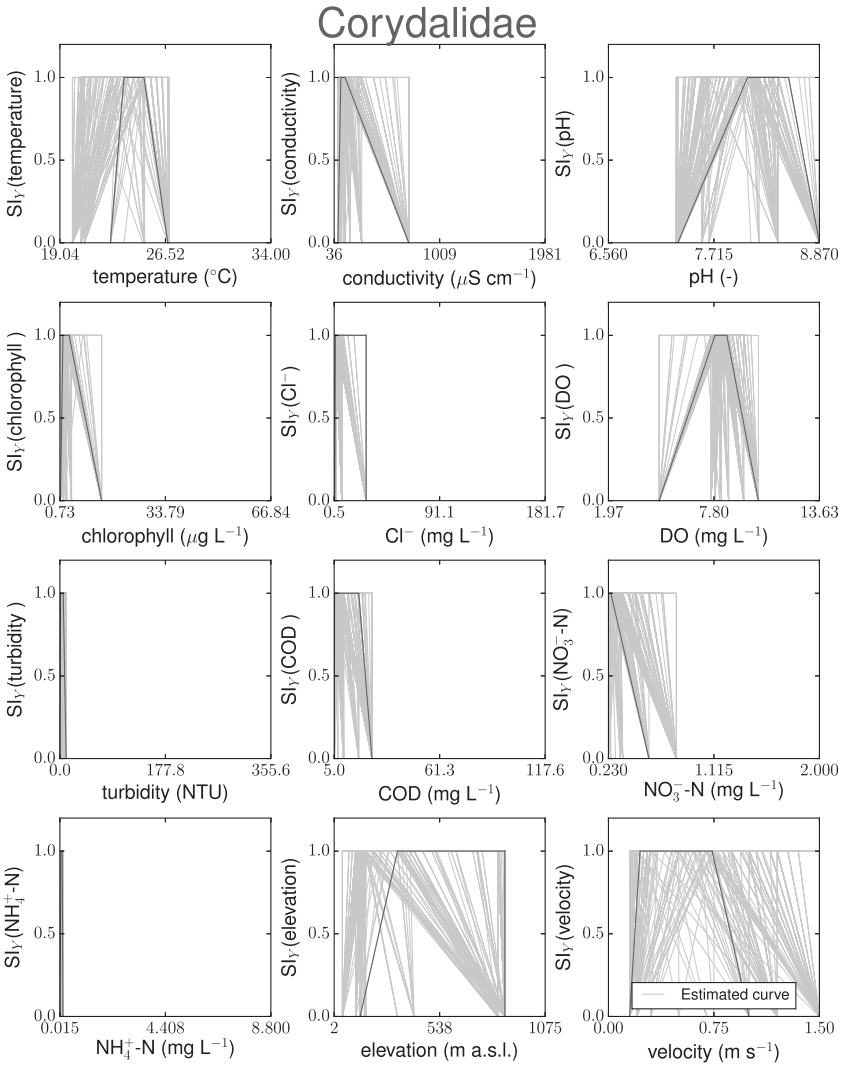

Fig. B6. Species response curves for the taxon Corydalidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

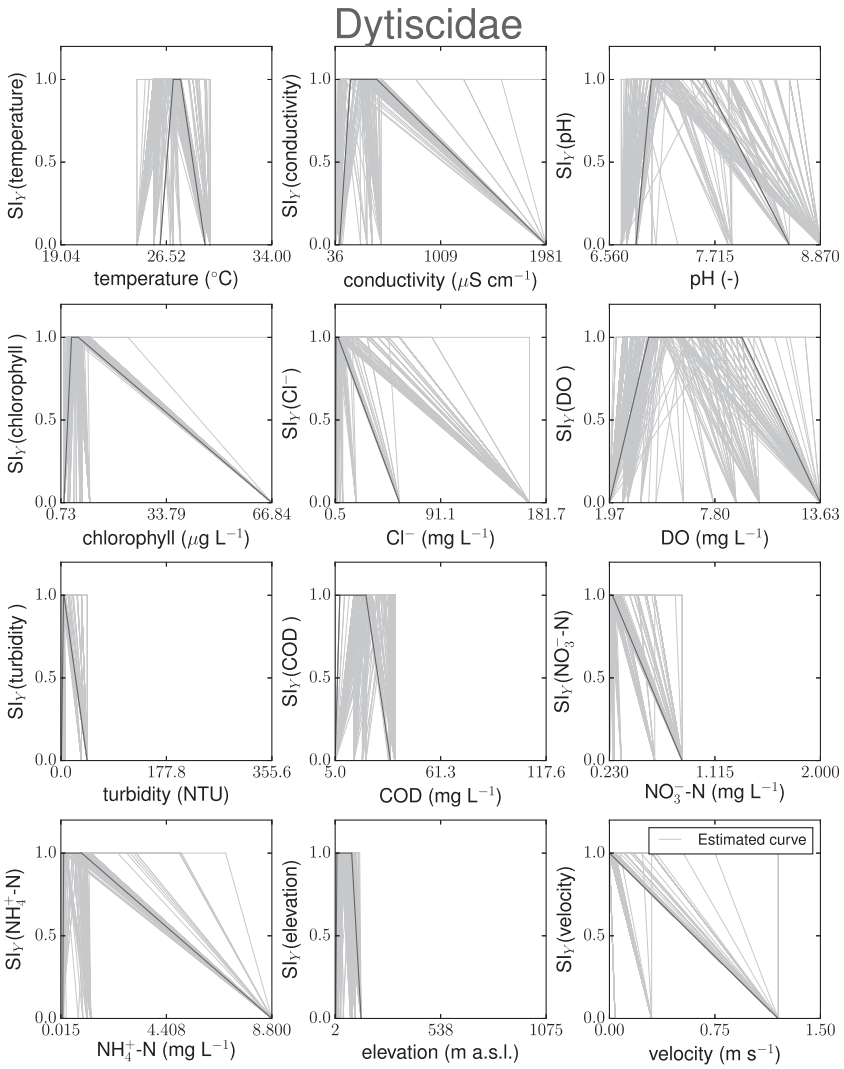

Fig. B7. Species response curves for the taxon Dytiscidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

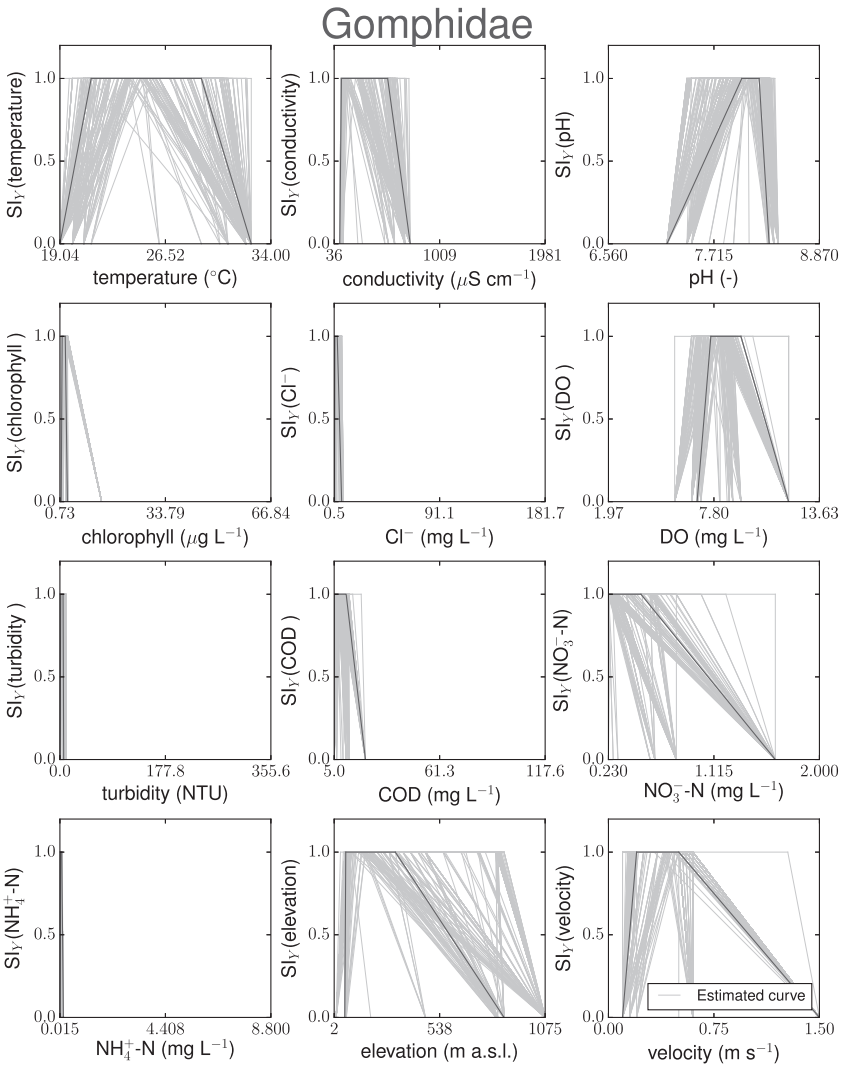

Fig. B8. Species response curves for the taxon Gomphidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

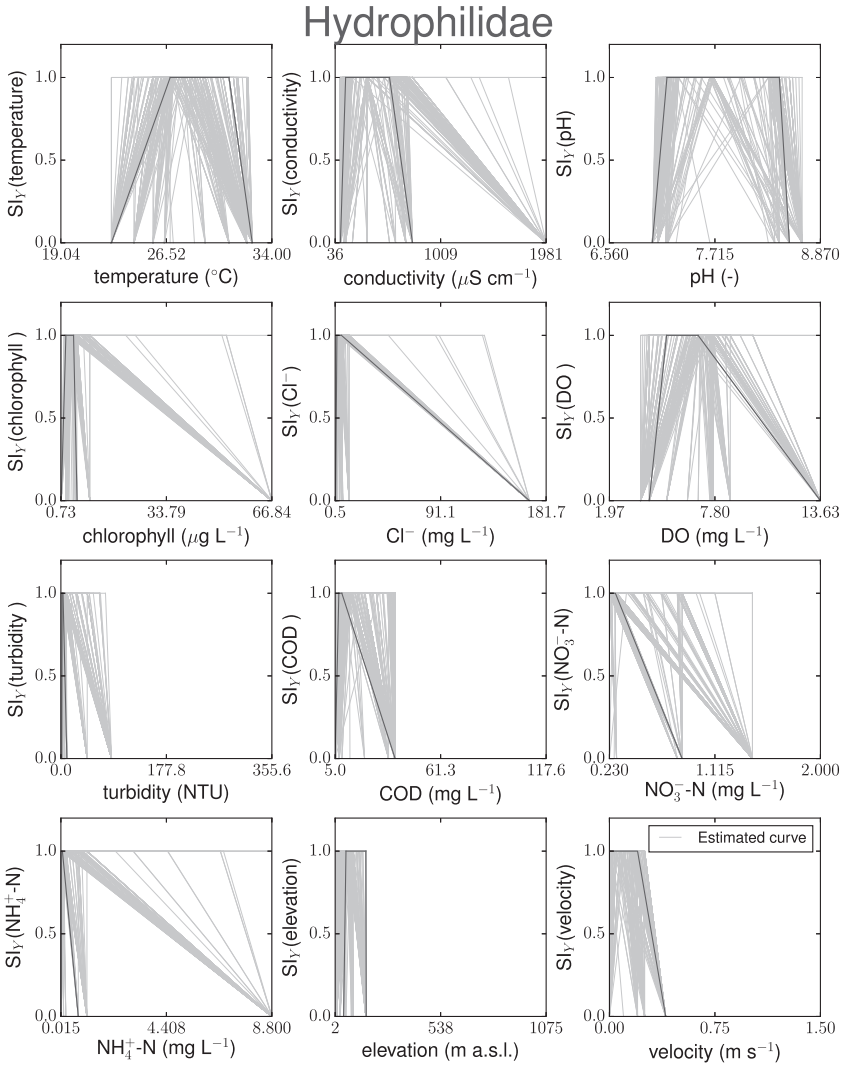

Fig. B9. Species response curves for the taxon Hydrophilidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

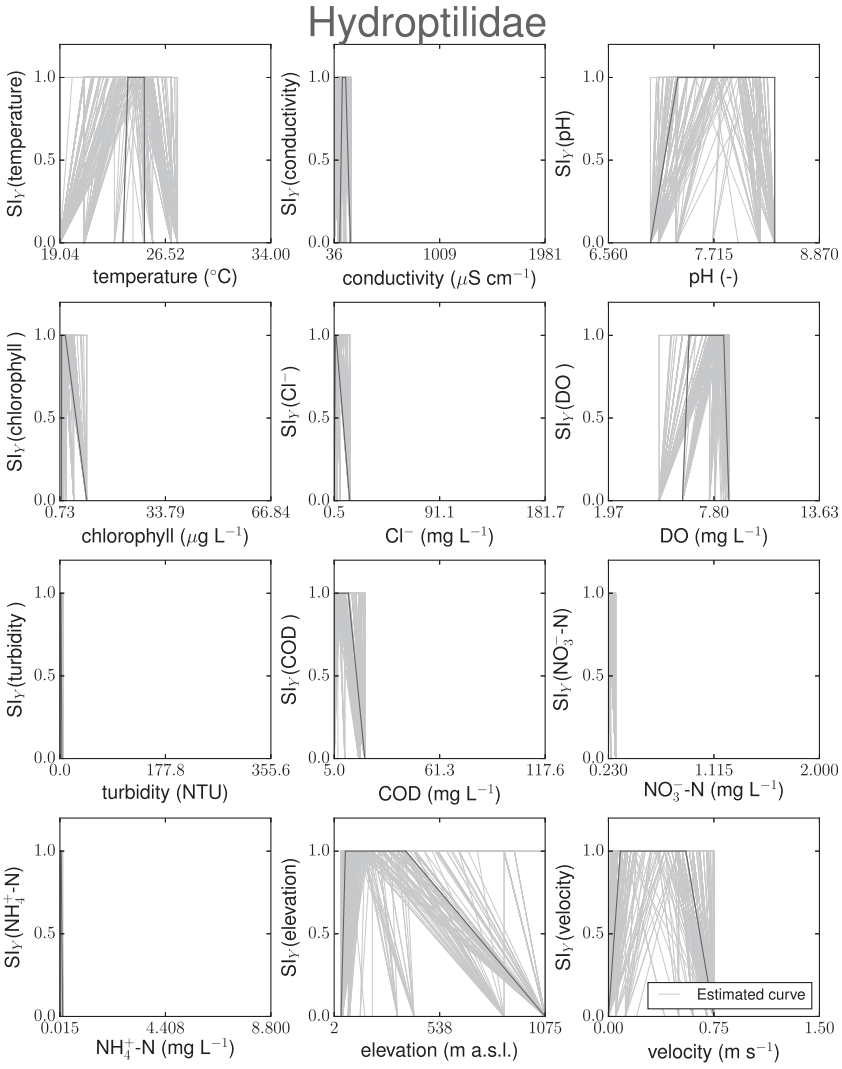

Fig. B10. Species response curves for the taxon Hydroptilidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

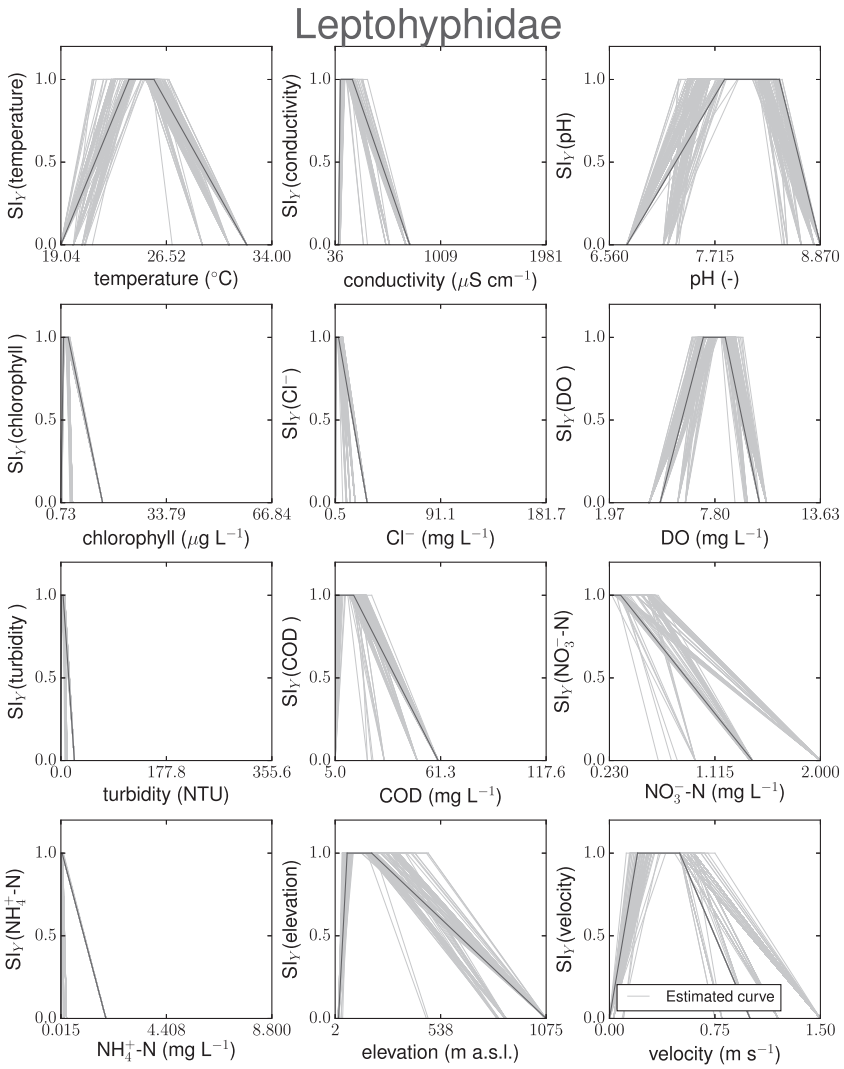

Fig. B11. Species response curves for the taxon Leptohyphidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

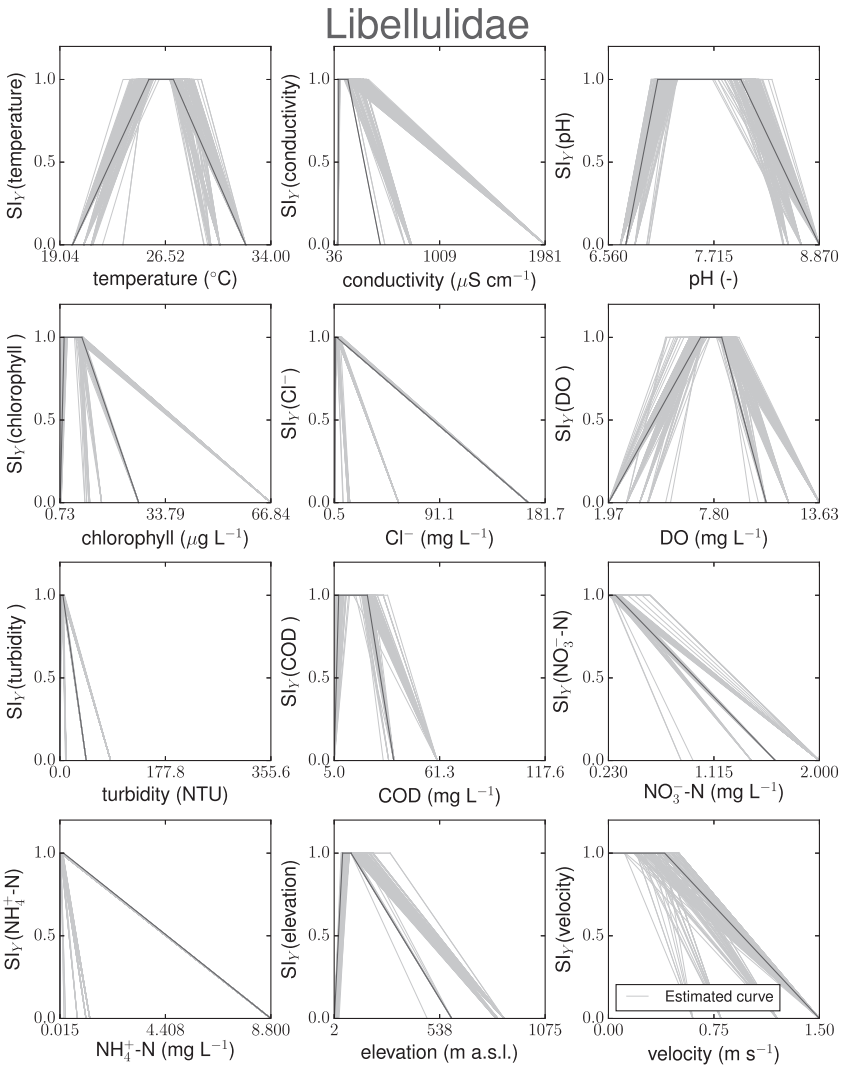

Fig. B12. Species response curves for the taxon Libellulidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

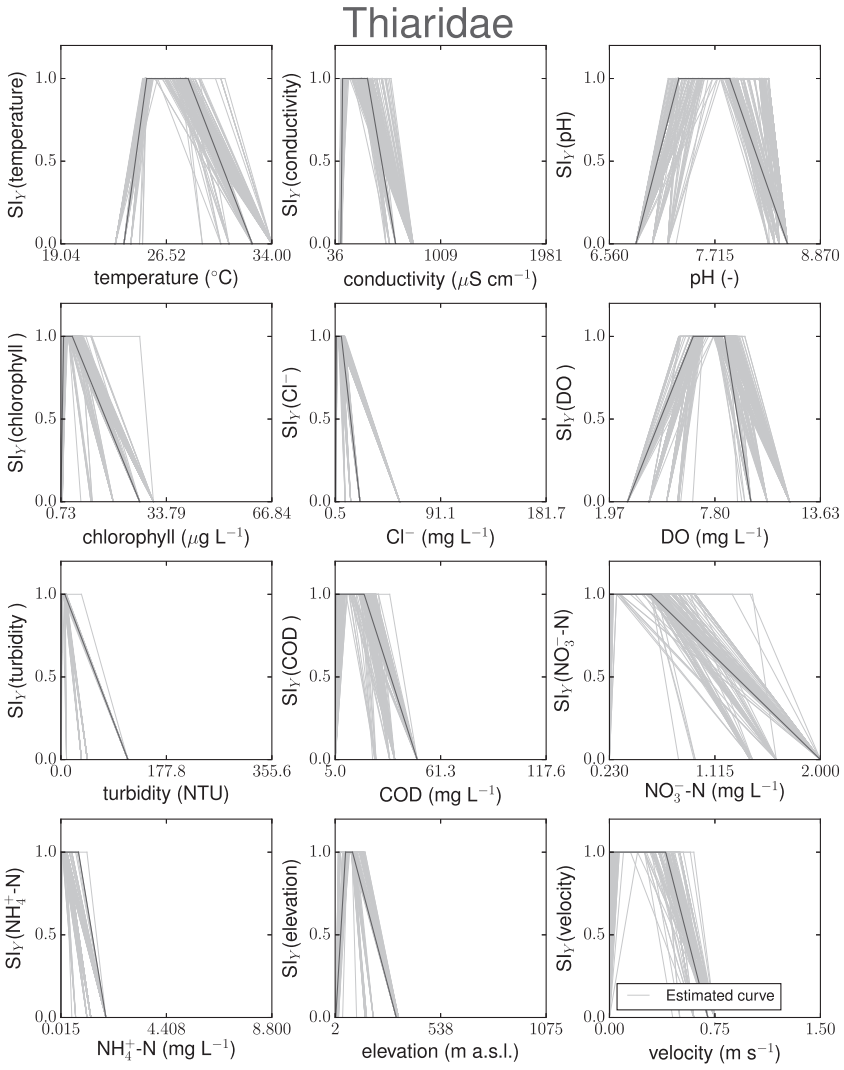

Fig. B13. Species response curves for the taxon Thiaridae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

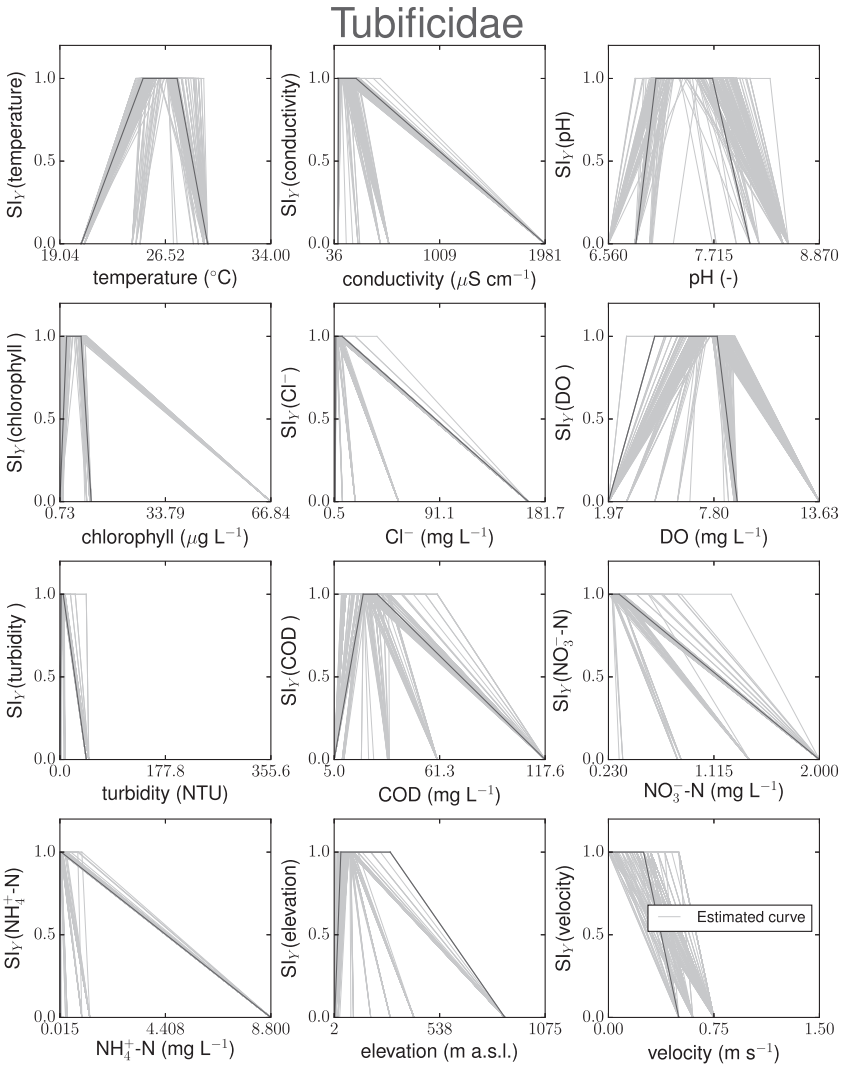

Fig. B14. Species response curves for the taxon Tubificidae for 12 variables. Different curves are obtained by estimating the response curve by resampling the original data. The species response curve in black is obtained by calculating the mean for the parameters a_1, a_2, a_3 and a_4 over all curves.

Appendix C. Algorithm performance

The values for the hyper parameters of the genetic algorithm are chosen with the guidelines of Gibbs et al. (2008). In the paper of Gibbs et al. (2008), a calibration methodology to determine the values for the hyper parameters, without the need of a trial-and-error approach is presented:

1 Determine the number of Function Evaluations (*FE*). In this paper, 4000 is used as value of *FE*, as this is the number of simulations needed to perform a grid search (12 variable, 2 options $= 2^{12}$-1 $= 4095$).
2 Solve equation C1 to find *R*, the population size, with $M = 3$ and $l = 12$ (length of binary string):

$$\frac{FE}{R}\log\left(1-\frac{1}{R}\right) = -M - \log\left(\sqrt{\frac{l}{12}}\right) \quad \text{(C1)}$$

3 Compute the mutation rate by dividing 5 by *R* (* 100 to obtain percentage).
4 Use elitism and a crossover rate of 100%.

With this methodology, the number of chromosomes is set to 24, the mutation rate to 20%, the crossover rate to 100%, the selection rate to 50%. In addition elitism is used. In order to account for the specific situation, a limited iterative approach is followed, by iterating over other values in the surrounding of the ones found with the guidelines of Gibbs et al. (2008). It was concluded that the acquired values were near-optimal. In order to test the efficiency of the SGA, the approach is tested to a grid search approach on 24-core Intel E5-2680v3, Haswell-EP @ 2.5 GHz computer node. The value of the objective function and the runtime is monitored as a function of the generation. In order to account for the effect of the initial conditions on the repeatability, the SGA is repeatedly run on the same data sample (for the species Baetidae).

In Figure C1, the value for the objective function (AIC) and the runtime is shown for the generation (lower panel). The standard deviation on the objective function and runtime is calculated as a measure for the uncertainty. In the upper panel, the percentage of algorithm runs that converged to the optimal solution is shown. On average, the objective function converges (±220 (50) to 260 (60) seconds (generations)) to a near-optimal solution before the grid search has completed (337, equivalent to 77 generations). It is important to note that for a number of simulations (4%), an near-optimal solution is reached after 50 generations, instead of the optimal solution. After 60 generations, 2% have reached a near-optimal solution, whereas after 77 generations, 1% have to converge to the optimal solution. After 100 generations, all simulations converged to the optimal solution. In general, the algorithm finds the optimal solution before the grid search approach, however, in some cases (1% of the simulations) a near-optimal solution is found.

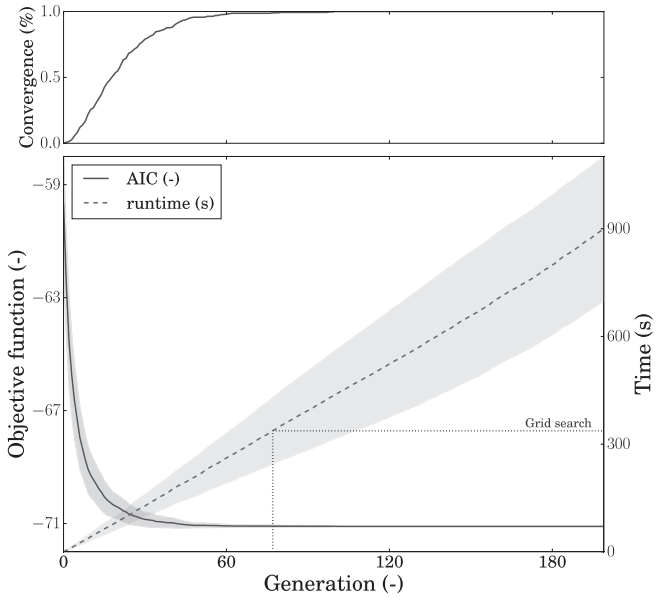

Fig. C1. Overview of performance of the SGA. In the upper pannel, the number of SGA analysis (%) that converged to the optimal solution for the number of generations is found. In the lower pannel, the evolution of the objective function and the runtime of the SGA is given as a function of the generation. The experiment is based on 300 repeated runs and compared to a grid search approach.

Appendix D. Stability

The SGA analysis is repeatedly run with different samples of the data, in order to estimate the effect of uncertainty in the ecological data. The required number of simulations is estimated by checking convergence of the support for the inclusion of a variable. In this approach, the support for the inclusion of a variable is calculated by dividing the number of SGA analysis which selected the variable as explanatory input variable by the total number of SGA analysis. By investigating the support over 10, 20, 30 and so on repeatedly run analysis, one can estimate the number where the statistics converge. In Figure D1 to D14, the stability of the support as a function of the number of SGA analysis is found for all taxa. From these figures, it can be visually observed that the statistics for the analyses converge after approximately 100–200 analysis. In this paper, 300 simulations are used.

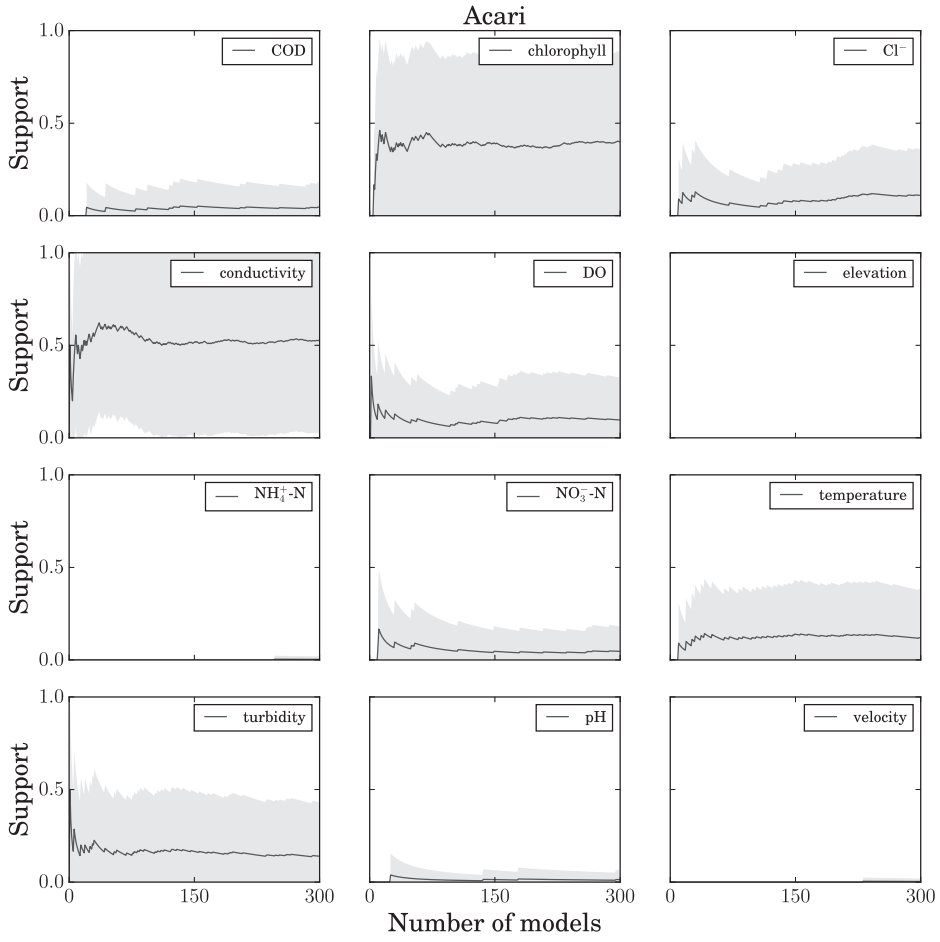

Fig. D1. Stability functions for Acari. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

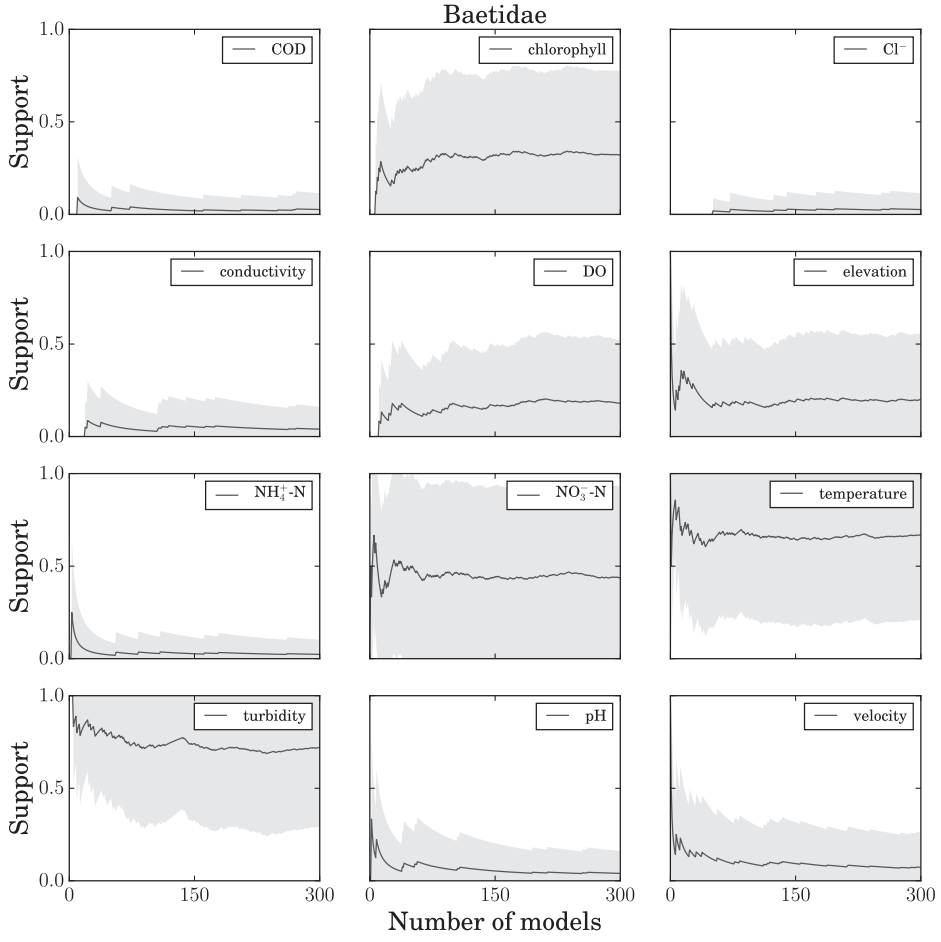

Fig. D2. Stability functions for Baetidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

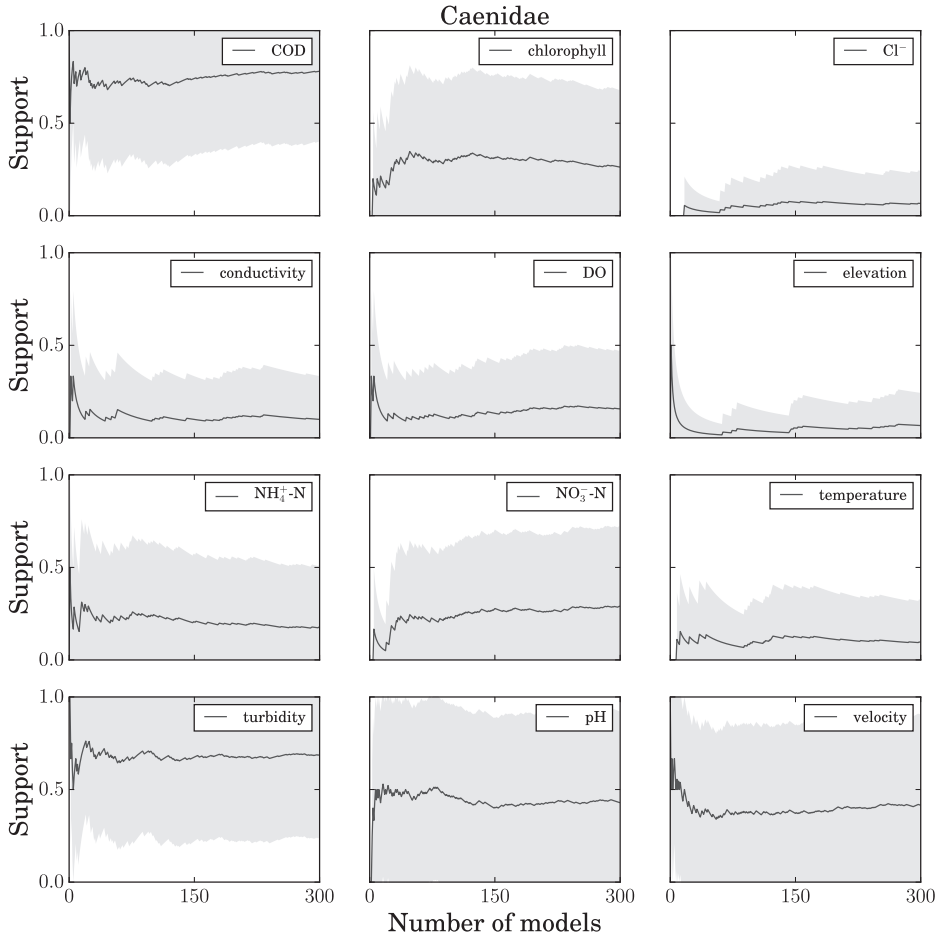

Fig. D3. Stability functions for Caenidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

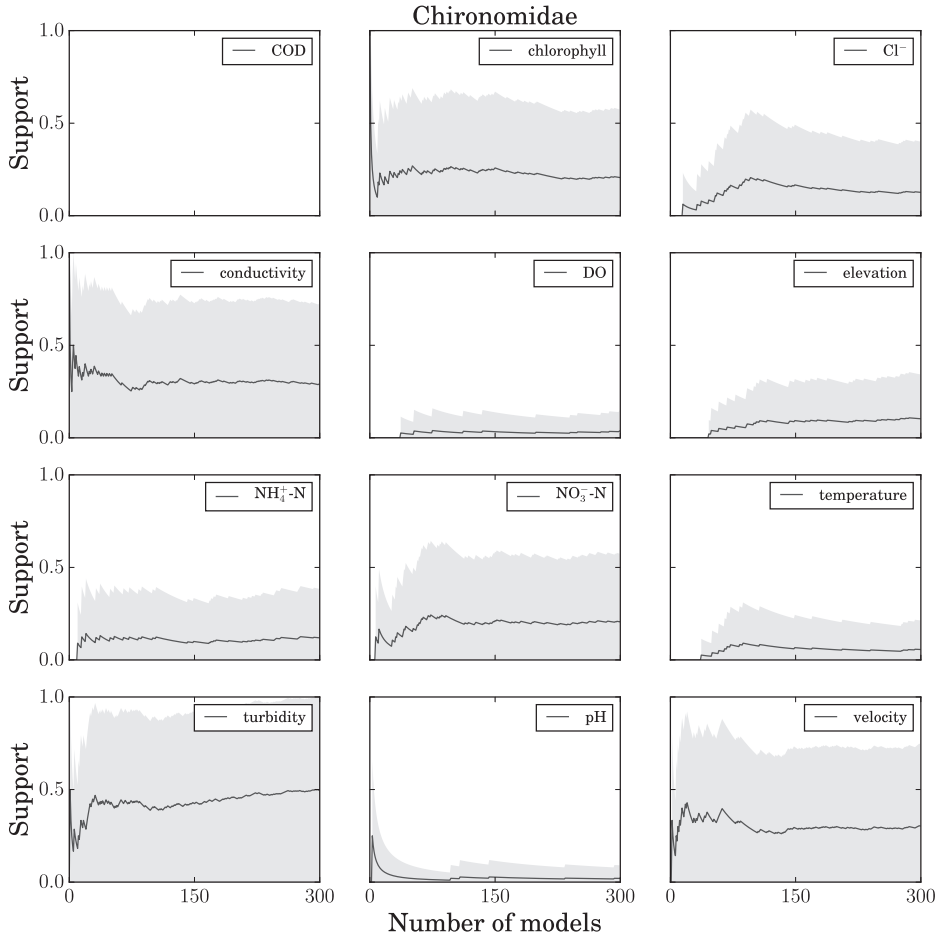

Fig. D4. Stability functions for Chironomidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

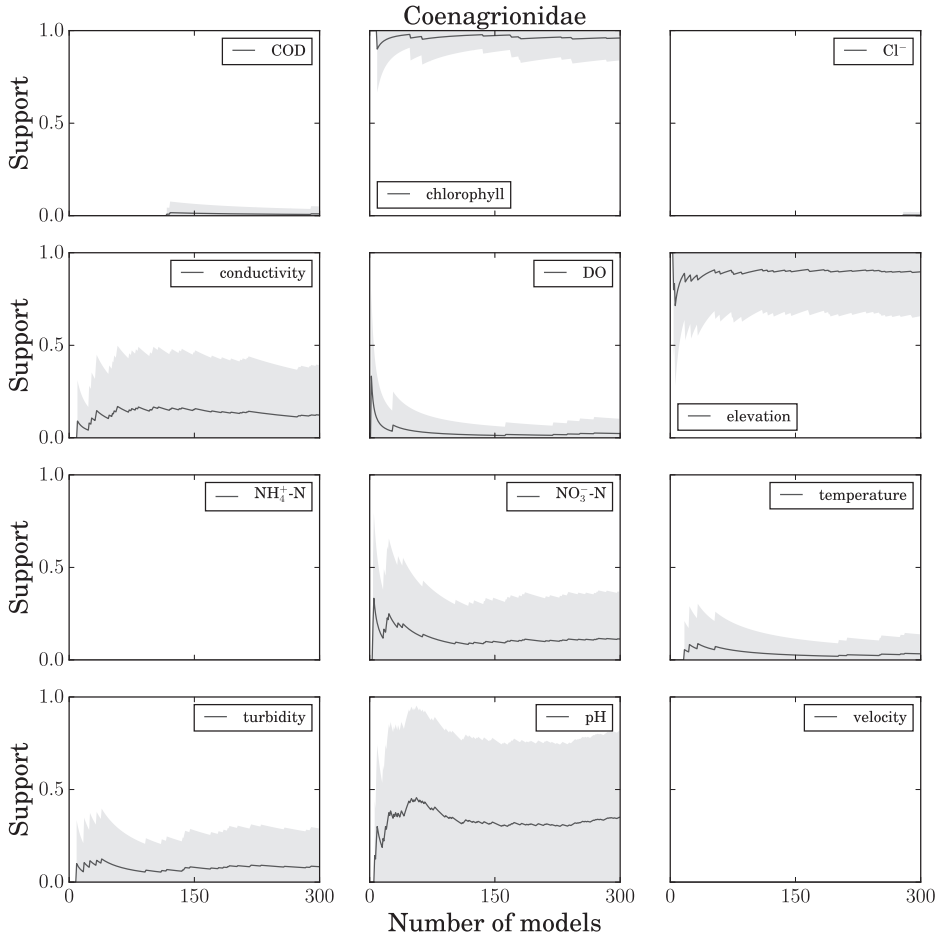

Fig. D5. Stability functions for Coenagrionidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

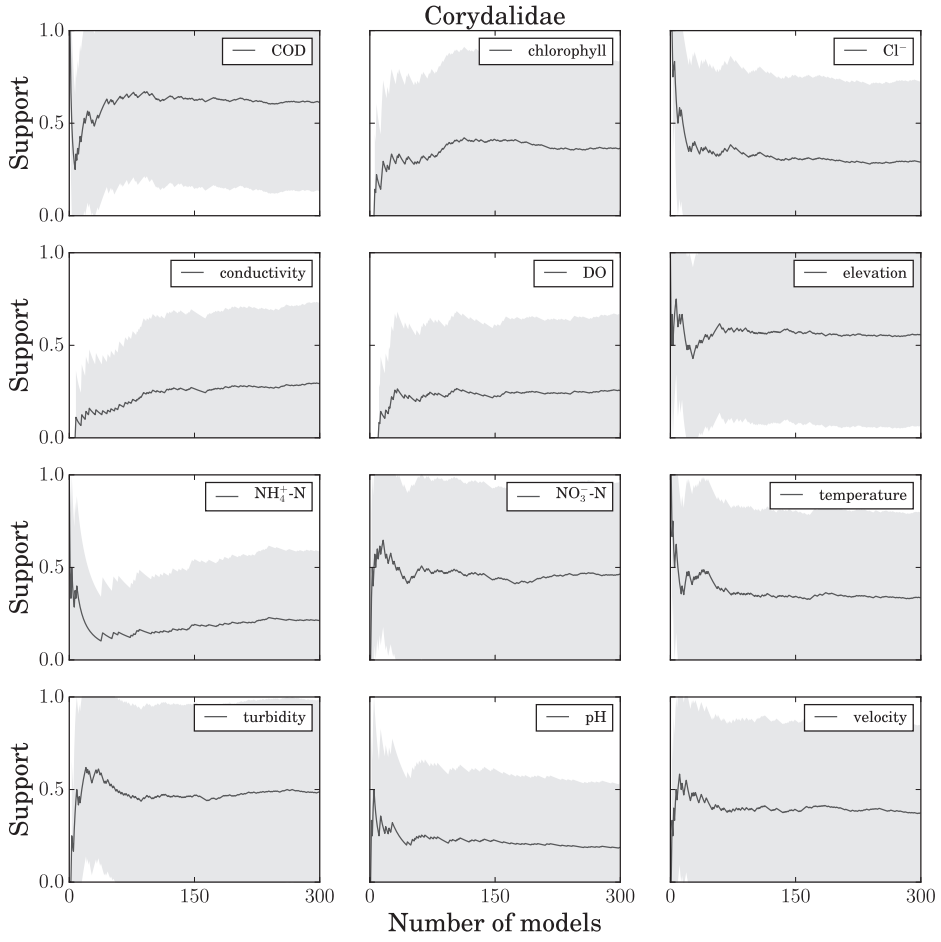

Fig. D6. Stability functions for Corydalidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

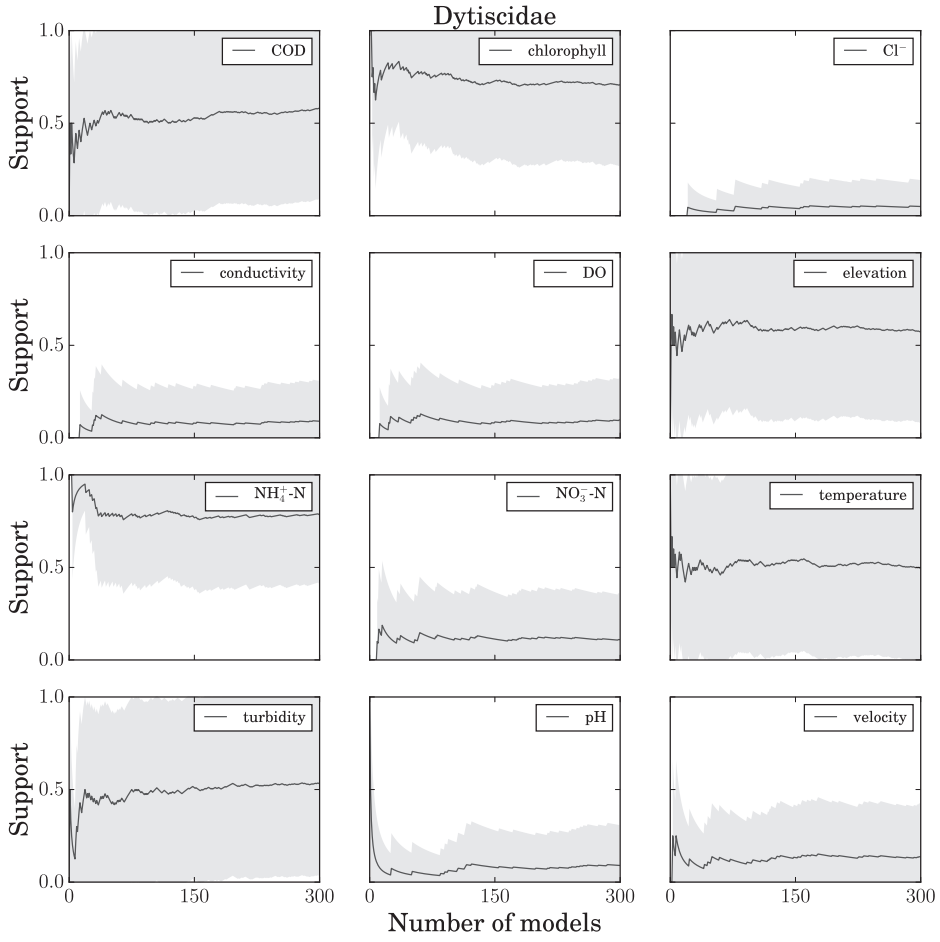

Fig. D7. Stability functions for Dytiscidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

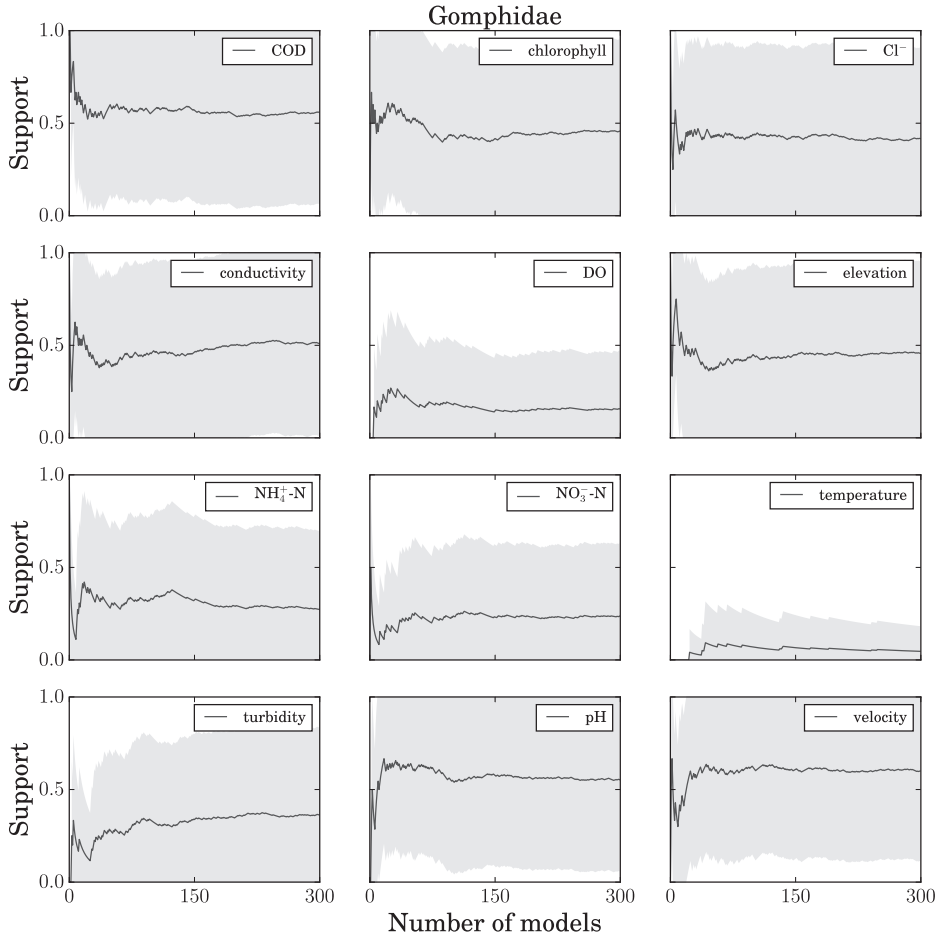

Fig. D8. Stability functions for Gomphidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

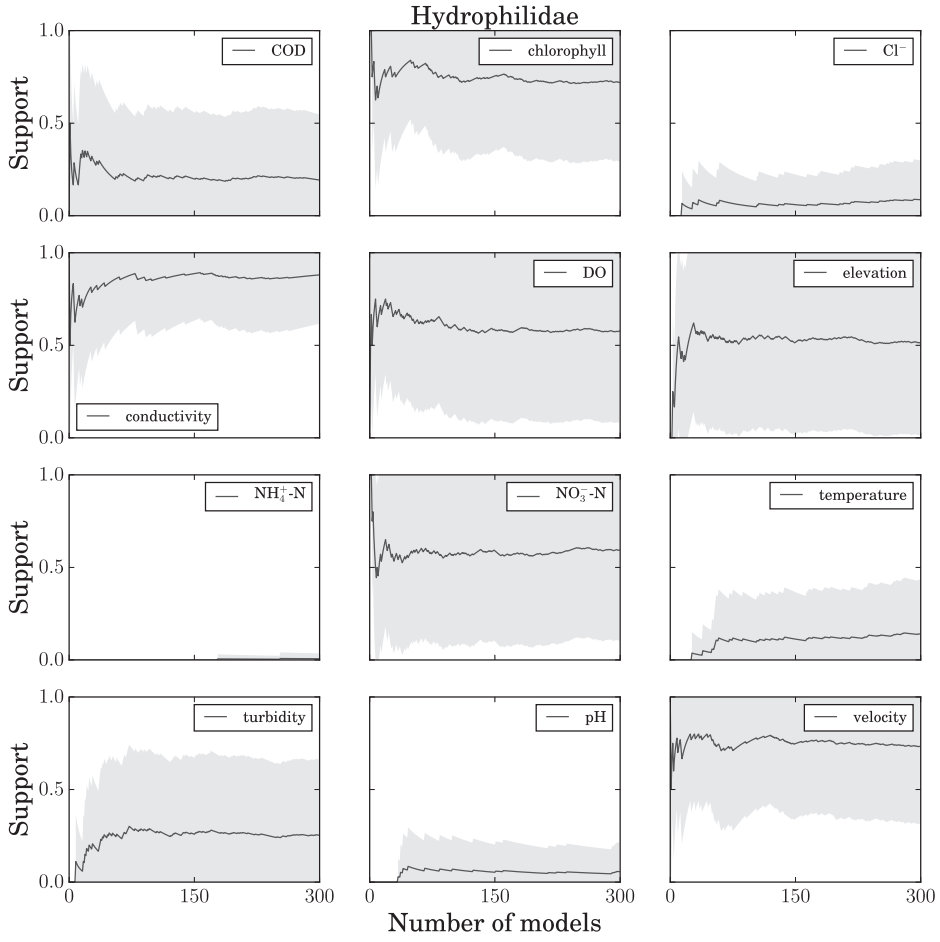

Fig. D9. Stability functions for Hydrophilidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

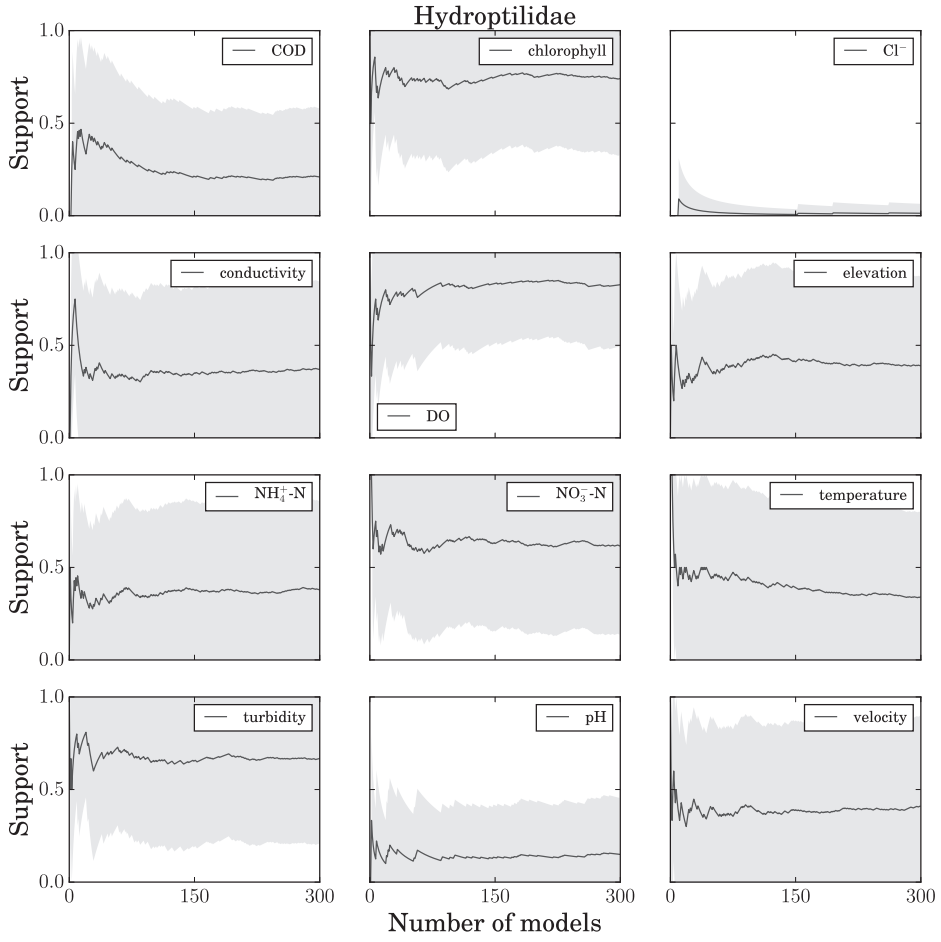

Fig. D10. Stability functions for Hydroptilidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

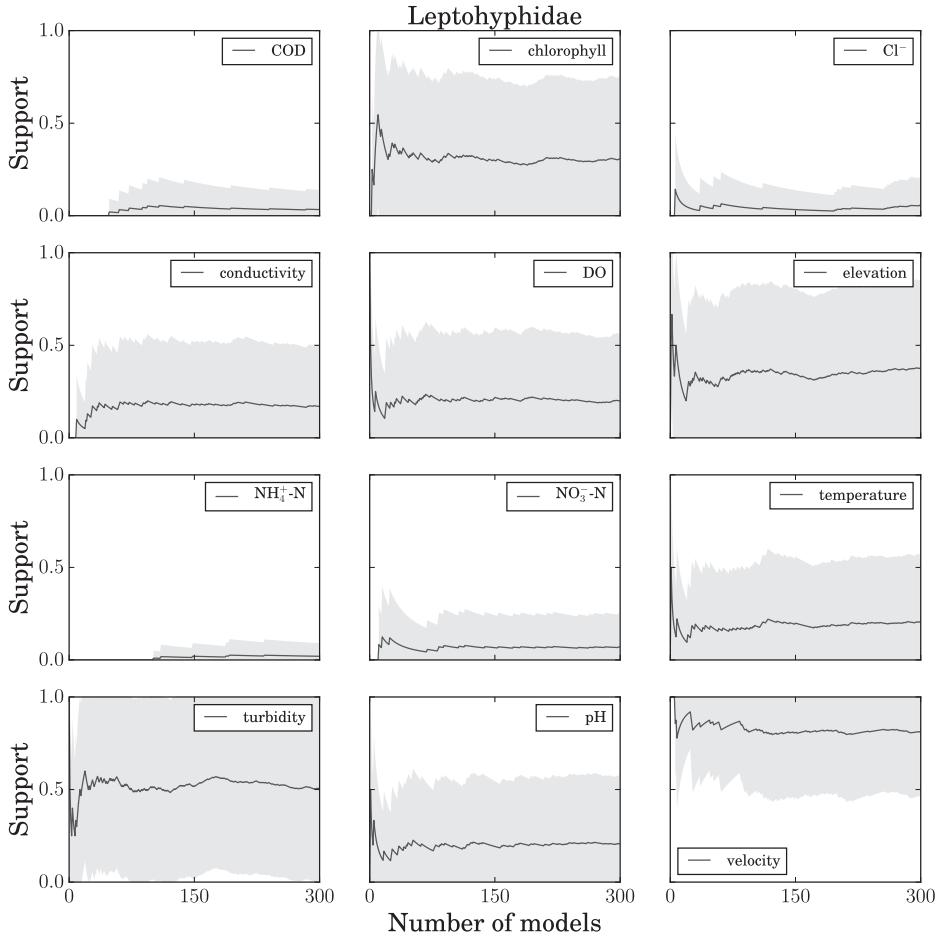

Fig. D11. Stability functions for Leptohyphidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

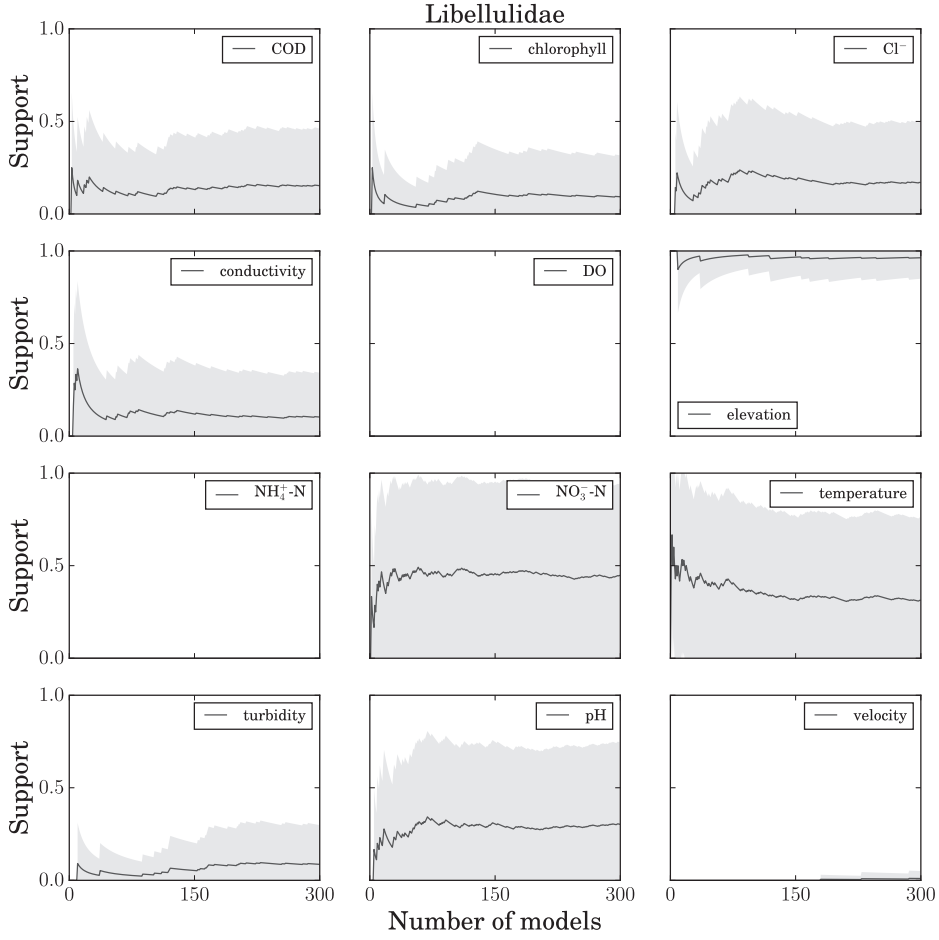

Fig. D12. Stability functions for Libellulidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

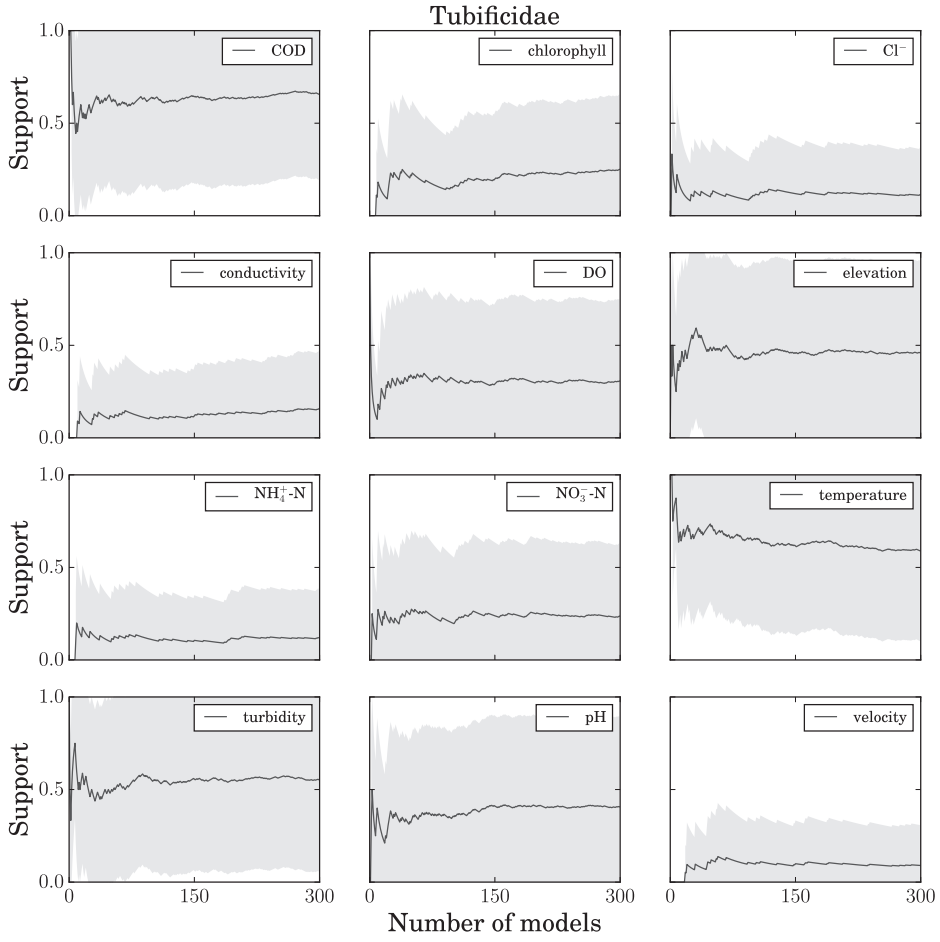

Fig. D13. Stability functions for Tubificidae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

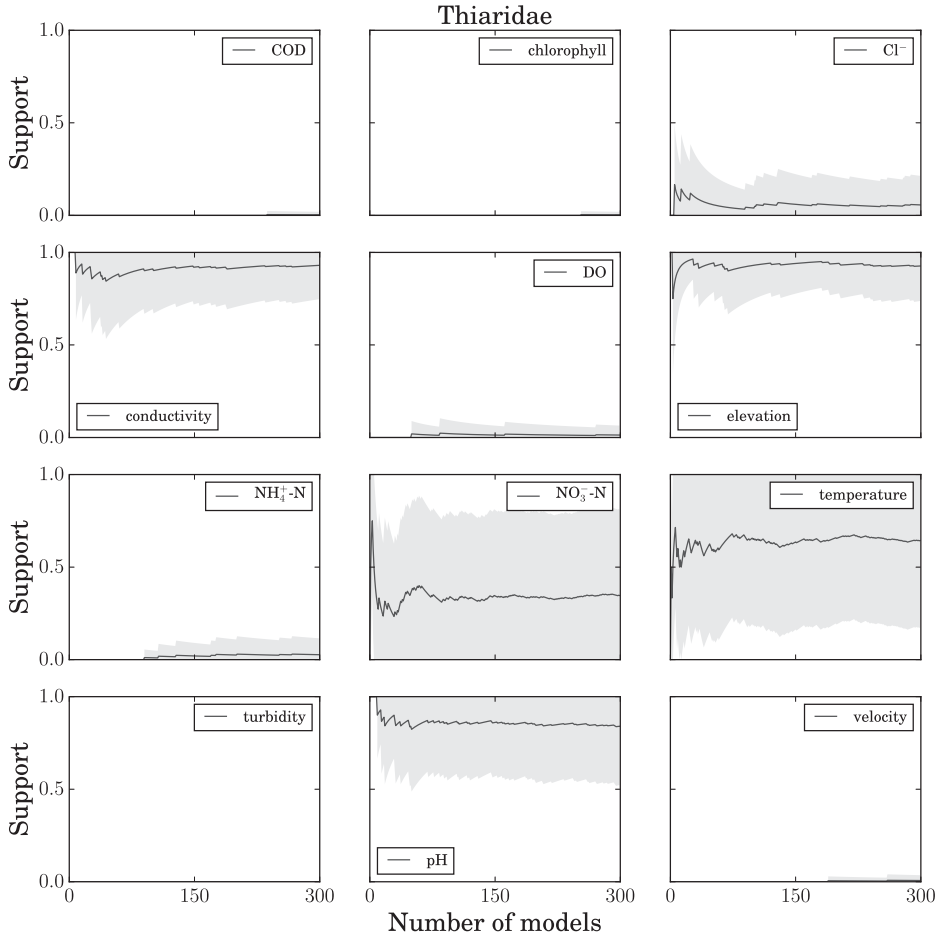

Fig. D14. Stability functions for Thiaridae. On the x-axis, the number of models found with a repeated SGA analysis is found. The SGA analysis is repeated with different data samples used for model construction and identificaiton. On the y-axis, the support, defined as number of SGA analysis in which the variable is present in the (near-)optimal solution divided by the total number of SGA analysis. The support is given in the black line, the uncertainty on this support (grey) is given by the Shannon entropy.

References

Adriaenssens, V., Goethals, P.L.M., De Pauw, N., 2006. Fuzzy knowledge-based models for prediction of Asellus and Gammarus in watercourses in Flanders (Belgium). Ecol. Model. 195, 3–10.

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

Ambarita, M.N.D., Lock, K., Boets, P., Everaert, G., Thi, H.T.N., Forio, M.A.E., Musonge, P.L.S., Semjonova, N., Bennetsen, E., Landuyt, D., Dominguez-Granda, L., Goethals, P.L.M., 2016. Ecological water quality analysis of the Guayas river basin (Ecuador) based on macroinvertebrates indices. Limnologica - Ecol. Manag. Inland Waters 27–59.

Ambelu, A., Lock, K., Goethals, P.L.M., 2010. Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. Ecol. Inf. 5, 147–152.

Arias-Hidalgo, M., Villa-Cox, G., Griensven, A.V., Solórzano, G., Villa-Cox, R., Mynett, A.E., Debels, P., 2013. A decision framework for wetland management in a river basin context: the “Abras de Mantequilla” case study in the Guayas River Basin, Ecuador. Environ. Sci. Policy 34, 103–114.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118.

Austin, M.P., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19.

Beale, C.M., Lennon, J.J., 2012. Incorporating uncertainty in predictive species distribution modelling. Philosophical Trans. R. Soc. B Biol. Sci. 367, 247–258.

Bennetsen, E., Gobeyn, S., Goethals, P.L.M., 2016. Species distribution models grounded in ecological theory for decision support in river management. Ecol. Model. 325, 1–12.

Boets, P., Holguin, G., Lock, K., Goethals, P.L.M., 2013. Data-driven habitat analysis of the Ponto-Caspian amphipod Dikerogammarus villosus in two invaded regions in Europe. Ecol. Inf. 17, 36–45.

Boets, P., Landuyt, D., Everaert, G., Broekx, S., Goethals, P.L.M., 2015. Evaluation and comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the habitat suitability for alien macroinvertebrates. Environ. Model. Softw. 74, 92–103.

Booth, T.H., Nix, H.A., Busby, J.R., Hutchinson, M.F., 2014. Bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Divers. Distributions 20, 1–9.

Boulangeat, I., Gravel, D., Thuiller, W., 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecol. Lett. 15, 584–593.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.

Burnham, K., Anderson, D., 2002. Model Selection and Multimodel Inference: a Practical Information-theoretic Approach, , 2 ed.vol. 172. Springer, New York.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodivers. Conservation 2, 667–680.

De Pauw, N., Vanhooren, G., 1983. Method for biological quality assessment of watercourses in Belgium. Hydrobiologia 100, 153–168.

De Pauw, N., Vannevel, R., 1991. Macroinvertebraten en waterkwaliteit. Determineersleutels van macroin-vertebraten en beoordelingsmethoden van de waterkwaliteit, 2 ed. Stichting Leefmilieu, Antwerpen.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol. Model. 160, 291–300.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol. Model. 195, 20–29.

Domínguez, E., Fernández, H.R., 2009. Macroinvertebrados Bentónicos Sudamericanos. Sistemática y Biología. Fundación Miguel Lillo, Tucumán, Argentina.

Domínguez-Domínguez, O., Martínez-Meyer, E., Zambrano, L., De León, G.P.P., 2006.

Using ecological-niche modeling as a conservation tool for freshwater species: live-bearing fishes in central Mexico. Conserv. Biol. 20, 1730–1739.
Domisch, S., Araújo, M.B., Bonada, N., Pauls, S.U., Jähnig, S.C., Haase, P., 2013. Modelling distribution in European stream macroinvertebrates under future climates. Glob. Change Biol. 19, 752–762.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.
Eiben, A.E., Smith, J., 2015. From evolutionary computation to the evolution of things. Nature 521, 476–482.
Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J., Peterson, A., Phillips, S., Richardson, K., Scachetti-Pereira, R., Schapire, R., Soberon, J., Williams, S., Wisz, M., Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Ellison, A.M., 2004. Bayesian inference in ecology. Ecol. Lett. 7, 509–520.
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.
Everaert, G., Pauwels, I.S., Boets, P., Buysschaert, F., Goethals, P.L.M., 2013. Development and assessment of ecological models in the context of the European Water Framework Directive: key issues for trainers in data-driven modeling approaches. Ecol. Inf. 17, 111–116.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.
Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2014. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol. Evol. 6, 424–438.
Forio, M.A.E., Landuyt, D., Bennetsen, E., Lock, K., Nguyen, T.H.T., Ambarita, M.N.D., Musonge, P.L.S., Boets, P., Everaert, G., Dominguez-Granda, L., Goethals, P.L.M., 2015. Bayesian belief network models to analyse and predict ecological water quality in rivers. Ecol. Model. 312, 222–238.
Fu, B., Guillaume, J.H.A., 2014. Assessing certainty and uncertainty in riparian habitat suitability models by identifying parameters with extreme outputs. Environ. Model. Softw. 60, 277–289.
Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., Mouton, A.M., 2013. Habitat prediction and knowledge extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species distribution models. Environ. Model. Softw. 47, 1–6.
Fukuda, S., Mouton, A.M., De Baets, B., 2012. Abundance versus presence/absence data for modelling fish habitat preference with a genetic Takagi-Sugeno fuzzy system. Environ. Monit. Assess. 184, 6159–6171.
Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. Aquat. Ecol. 41, 427–441.
Gabriels, W., Lock, K., De Pauw, N., Goethals, P.L.M., 2010. Multimetric macroinvertebrate index flanders (MMIF) for biological assessment of rivers and lakes in flanders (Belgium). Limnologica 40, 199–207.
Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51.
Gibbs, M.S., Dandy, G.C., Maier, H.R., 2008. A genetic algorithm calibration method based on convergence due to genetic drift. Inf. Sci. 178, 2857–2869.
Gibbs, M.S., Maier, H.R., Dandy, G.C., 2015. Using characteristics of the optimisation problem to determine the Genetic Algorithm population size when the number of evaluations is limited. Environ. Model. Softw. 69, 226–239.
Gies, M., Sondermann, M., Hering, D., Feld, C.K., 2015. A comparison of modelled and actual distributions of eleven benthic macroinvertebrate species in a Central European mountain catchment. Hydrobiologia 758, 123–140.
Gobeyn, S., Bennetsen, E., Van Echelpoel, W., Everaert, G., Goethals, P.L.M., 2016. Impact of abundance data errors on the uncertainty of an ecological water quality assessment index. Ecol. Indic. 60, 746–753.
Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston.
Goldberg, D.E., Deb, K., 1991. A comparative analysis of selection schemes used in genetic algorithms. Found. Genet. Algorithms 1, 69–93.
Grinnell, J., 1917. The niche-relationships of the California thrasher. Auk 34, 427–433.
Guisan, A., Rahbek, C., 2011. SESAM a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. J. Biogeogr. 38, 1433–1444.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Hamblin, S., 2013. On the practical usage of genetic algorithms in ecology and evolution. Methods Ecol. Evol. 4, 184–194.
Heikkinen, J., Mäkipää, R., 2010. Testing hypotheses on shape and distribution of ecological response curves. Ecol. Model. 221, 388–399.
Hernández, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.
Hirzel, A.H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. J. Appl. Ecol. 45, 1372–1381.
Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods Ecol. Evol. 5, 506–513.
Hutchinson, E.G., 1957. Concluding remarks. Cold Spring Harb. Symposia Quantative Biol. 159, 415–427.
Karl, J., Svancara, L., Heglund, P., Wright, N., Scott, J., 2002. Species commonness and the accuracy of habitatrelationships models. chapter Species co. In: Scott, J., Heglund, P., Morrison, M., Haufler, J., Raphael, M., Wall, W., Samson, F. (Eds.), Predicting Species Occurrences. Issues of Accuracy and Scale. Island Press, Washington, pp. 573–580.
Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.C., Zimmermann, N.E., O'Hara, R.B., 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. J. Biogeogr. 39, 2163–2178.
Langhans, S.D., Reichert, P., Schuwirth, N., 2014. The method matters: a guide for indicator aggregation in ecological assessments. Ecol. Indic. 45, 494–507.
Lock, K., Adriaens, T., Goethals, P., 2014. Effect of water quality on blackflies (Diptera: Simuliidae) in Flanders (Belgium). Limnologica 44, 58–65.
Maier, H.R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L.S., Cunha, M.C., Dandy, G.C., Gibbs, M.S., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D.P., Vrugt, J.A., Zecchin, A.C., Minsker, B.S., Barbour, E.J., Kuczera, G., Pasha, F., Castelletti, A., Giuliani, M., Reed, P.M., 2014. Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. Environ. Model. Softw. 62, 271–299.
Manel, S., Ceri Williams, H., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.
May, R., Dandy, G., Maier, H., 2011. Review of input variable selection methods for artificial neural networks. August 2016. In: Artificial Neural Networks - Methodological Advances and Biomedical Applications, p. 362.
McPherson, J.M., Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. Ecography 30, 135–151.
Mount, N., Maier, H., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.J., Abrahart, R., 2016. Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the panta rhei science plan. Hydrological Sci. J. 61, 1192–1208.
Mouton, A.M., Alcaraz-Hernández, J.D., De Baets, B., Goethals, P.L.M., Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environ. Model. Softw. 26, 615–622.
Mouton, A.M., De Baets, B., Goethals, P.L.M., 2009. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. Environ. Model. Softw. 24, 982–993.
Mouton, A.M., De Baets, B., Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecol. Model. 221, 1995–2002.
Muñoz-Mas, R., Fukuda, S., Vezza, P., Martínez-Capel, F., 2016. Comparing four methods for decision-tree induction: a case study on the invasive Iberian gudgeon (Gobio lozanoi; Doadrio and Madeira, 2004). Ecol. Inf. 34, 22–34.
Poff, N.L., 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. J. north Am. Benthol. Soc. 16, 391–409.
Raleigh, R.B., Zuckerman, L.D., Nelson, P.C., 1986. Habitat suitability index models and instream flow suitability curves: brown trout. U.S. Fish Wildl. Serv. 82, 65.
Reed, P., Minsker, B., Goldberg, D.E., 2000. Designing a competent simple genetic algorithm for search and optimization. Water Resour. Res. 36, 3757–3761.
Sadeghia, R., Zarkami, R., Sabetraftar, K., Van Damme, P., 2013. Application of genetic algorithm and greedy stepwise to select input variables in classification tree models for the prediction of habitat requirements of Azolla filiculoides (Lam.) in Anzali wetland. Iran. Ecol. Model. 251, 44–53.
Schmidt-Kloiber, A., Hering, D., 2015. An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. Ecol. Indic. 53, 271–282. www.freshwaterecology.info.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Soberón, J., Nakamura, M., 2009. Niches and distributional areas: concepts, methods, and assumptions. Proc. Natl. Academy Sci. U. S. A 106 (Suppl. l), 19644–19650.
Stockwell, D.R.B., Noble, I.R., 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math. Comput. Simul. 33, 385–390.
Termansen, M., McClean, C.J., Preston, C.D., 2006. The use of genetic algorithms and Bayesian classification to model species distributions. Ecol. Model. 192, 410–424.
Van Broekhoven, E., Adriaenssens, V., De Baets, B., Verdonschot, P.F., 2006. Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. Ecol. Model. 198, 71–84.
Vayghan, A.H., Zarkami, R., Sadeghi, R., Fazli, H., 2016. Modeling habitat preferences of caspian kutum, Rutilus frisii kutum (Kamensky, 1901) (Actinopterygii, cypriniformes) in the caspian sea. Hydrobiologia 766, 103–119.

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F., Mouton, A.M., 2015. Random forests to evaluate biotic interactions in fish distribution models. Environ. Model. Softw. 67, 173–183.
Waite, J.P., 1982. Competition for water resources of the rio Guayas, Ecuador. Optim. Allocation Water Resour. 135, 79–88.
Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P., Svenning, J.C., 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biol. Rev. 88, 15–30.
Zarkami, R., Sadeghi, R., Goethals, P.L.M., 2014. Modelling occurrence of roach "Rutilus rutilus" in streams. Aquat. Ecol. 48, 161–177.
Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods Ecol. Evol. 1, 3–14.