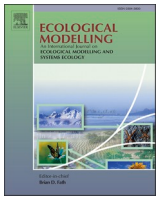
Ecological Modelling 481 (2023) 110353

Contents lists available at ScienceDirect

Ecological Modelling

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



A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation

Emma Chollet Ramampiandra [a,b,*], Andreas Scheidegger [a], Jonas Wydler [a,c], Nele Schuwirth [a,b]

[a] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600, Dübendorf, Switzerland
[b] ETH Zürich, Institute of Biogeochemistry and Pollutant Dynamics, 8092 Zürich, Switzerland
[c] University of Zürich, Department of Geography, 8057 Zürich, Switzerland

ARTICLE INFO

Keywords:
Species distribution model
Statistical models
Interpretable machine learning
Model complexity
Freshwater macroinvertebrates

ABSTRACT

Species distribution models are commonly applied to predict species responses to environmental conditions. A wide variety of models with different properties exist that vary in complexity, which affects their predictive performance and interpretability. Machine learning algorithms are increasingly used because they are capable to capture complex relationships and are often better in prediction. However, to inform environmental management, it is important that a model predicts well for the right reasons. It remains a challenge to select a model with a reasonable level of complexity that captures the true relationship between the response and explanatory variables as good as possible rather than fitting to the noise in the data.

In this study we ask: 1) how much predictive performance can we gain by using increasingly complex models, 2) how does model complexity affect the degree of overfitting, and 3) do the inferred responses differ among models and what can we learn from them? To address these questions, we applied eight models with different complexity to predict the probability of occurrence of freshwater macroinvertebrate taxa based on 2729 Swiss monitoring samples. We compared the models in terms of predictive performance during cross-validation and for generalization out of the calibration domain ("extrapolation" or transferability). We applied model agnostic tools to shed light on model interpretability.

Contrary to our expectation, all models predicted similarly well during cross-validation, while no model predicted better than the null model during out-of-domain generalization on average over all taxa. Performance was best for taxa with intermediate prevalence. More complex models predicted slightly better than standard statistical models but were prone to overfitting.

Overfitting indicates that a model describes not only the signal in the data but also part of the noise. This impedes the interpretation of response shapes learned by the model, because one cannot distinguish the signal from the noise. Furthermore, the strongly overfitting models learned irregular relationships and strong interactions that are ecologically not plausible. Thus, in this study, the minor gain in predictive performance from more complex models was outweighed by the overfitting.

Ecological field data that is used as model input or for calibration is typically prone to different sources of variability, from sampling, the measurement process and stochasticity. We therefore call for caution when using complex data-driven models to learn about species responses or to inform environmental management. In such cases, we recommend to compare a range of models regarding their predictive performance, overfitting and response shapes to better understand the robustness of inferred responses.

1. Introduction

A central question in ecology is to understand how species respond to environmental conditions. Species distribution models (SDMs) are useful tools to infer effects of environmental conditions on the distributions of organisms, i.e., to quantify their realized niches. They are also used to make predictions and inform environmental management (Linke et al., 2008; Araújo et al., 2019; Timoner et al., 2021), which are

Abbreviations: CV, Cross-Validation; ML, Machine Learning; ODG, Out-of-Domain Generalization; SDM, Species Distribution Model.

* Corresponding author.
E-mail address: emma.chollet@eawag.ch (E. Chollet Ramampiandra).

https://doi.org/10.1016/j.ecolmodel.2023.110353
Received 19 December 2022; Received in revised form 17 February 2023; Accepted 8 March 2023
Available online 3 April 2023

urgently needed in the current situation of climate change and biodiversity loss (IPBES, 2019). Whether to learn about species responses or to inform environmental management, we have to be confident that the models predict well for the right reasons, i.e., because they capture true relationships, and are not based on spurious correlations in the data (Schuwirth et al., 2019; Arif and MacNeil, 2022).

Many different statistical modeling approaches and machine learning (ML) algorithms are used for species distribution models (Elith and Franklin, 2013; Beery et al., 2021). The SDMs differ in their data requirements and their mathematical properties, which allow them to model non-linear response shapes, interactions between explanatory variables or multiple taxa at the same time. This results in various levels of complexity and ease of interpretability; simpler models impose a more constrained structure, usually easier to interpret, while more complex models or ML algorithms provide a more flexible structure, which can be perceived as black-box. The difference in the flexibility of the response shapes impacts the ability to fit to the calibration data and to make accurate predictions on unseen data. Overfitting arises, if the model allows too much flexibility in the response shapes or if the amount of data is limited and there is high variability in the data that cannot be explained by the influence factors, which is often the case in ecological datasets (Møller and Jennions, 2002; Barry and Elith, 2006). A model describing noise rather than species responses can be misleading in the context of decision support for environmental management. This is especially the case, when the model is projected in space or time, because overfitting reduces transferability (Randin et al., 2006), but also when the goal is to learn from the inferred response shapes within the domain of calibration. Therefore, overfitting should be quantified and the inferred response shapes should be investigated and assessed for plausibility.

In recent years, complex ML algorithms like random forest (RF) or artificial neural networks (ANN) have been increasingly applied because they have shown to perform well on big data sets in terms of predictive performance. However, they are harder to interpret due to their inaccessible internal structure (Rahman et al., 2021; Visser et al., 2022). Conversely, statistical models like generalized linear models (GLM) are easier to interpret, but have a less flexible structure and are therefore less able to capture complex patterns in the data (Guisan et al., 2002; Elith and Graham, 2009). Ensemble modeling (applying various models to the same dataset and averaging their output) can be a good strategy, if one is mainly interested in robust predictions (Araújo and New, 2007). However, ensemble modeling is not intended to be easily interpretable and therefore not very useful to learn about the system. In summary, more complex ML models are not necessarily superior in all cases and it remains challenging to select the right level of complexity for each application.

Many comparative studies have been conducted to determine which SDMs have the best predictive performance, within the domain of calibration (Elith and Graham, 2009; Li and Wang, 2013; Rahman et al., 2021; Stupariu et al., 2021; Visser et al., 2022) and when generalizing out of calibration domain, i.e., assessing model transferability (Tuanmu et al., 2011; Werkowska et al., 2017). Some also investigate what the different models learned by using model agnostic tools, i.e., tools that can be applied to any model, like variable importance assessment or visualization of response shapes (Zurell et al., 2012; Wenger and Olden, 2012; Fukuda et al., 2013; Molnar, 2019; Lucas, 2020). Various studies go a step further and interpret model responses to enhance our understanding about species niches or to support environmental management (Elith et al., 2008; Srivastava et al., 2019; Urbina-Cardona et al., 2019; Ryo et al., 2021). However, it remains unclear, if the best predicting models were prone to overfitting or not, which would affect the plausibility of the learned response shapes (Merow et al., 2014). To our knowledge, there have been few attempts to compare the learned response shapes of different models and systematically link them with model properties, performance and especially degree of overfitting.

The aim of this study is therefore to systematically assess the predictive performance, the degree of overfitting and the learned response shapes of SDM approaches with differing complexity. We ask: 1) how much predictive performance can we gain by using increasingly complex models, 2) how does model complexity affect the degree of overfitting, and 3) do the inferred responses differ among models and what can we learn from them? We apply eight statistical and machine learning models including different properties to a nationwide macroinvertebrate presence-absence biomonitoring data set from Swiss streams that spans 10 years of monitoring and 2729 observations. Macroinvertebrates are used as bioindicators to assess the ecological state of streams. There is therefore high interest in their response to different natural and human-induced environmental factors, like water quality, hydromorphology and temperature. We assess the predictive performance of the models during three-fold cross-validation (CV) and out-of-domain generalization (ODG). The latter is important to understand how useful the different models are to make predictions for changes that go beyond the conditions that are covered in the calibration data (e.g., highly relevant for climate change impact studies). Finally, we evaluate how the model properties and model complexity impact their interpretability.

Based on previous studies, we expect ML models to have higher predictive performance during CV compared to standard statistical models, because of the higher flexibility of the relationships between the response and explanatory variables (Fukuda et al., 2013; Li and Wang, 2013). We expect that especially complex interactions between explanatory variables can improve the predictive performance compared to simpler generalized linear models that do not include interaction terms. However, we expect ML models to be worse in ODG, because they have fewer mathematical constraints (Pearson et al., 2006; Wenger and Olden, 2012). Finally, by comparing the response shapes learned by the different models, we expect to determine model deficits of the simpler models and identify overfitting.

2. Material and methods

2.1. Macroinvertebrate dataset

We used presence-absence data of stream macroinvertebrate taxa from the Swiss Biodiversity Monitoring program (BDM) and other federal and cantonal monitoring programs. The dataset consists of 2729 observations taken at 1802 different sites covering the whole of Switzerland. The data was accessed via the MIDAT database, run by info fauna CSCF (Centre Suisse de Cartographie de la Faune) & karch. We selected data sampled between 2010 and 2019, where each site has been sampled between one and eight times by applying the same multihabitat sampling strategy (OFEV, 2019). For this study, we selected the 60 taxa that have a prevalence (i.e., the proportion of observations where the taxon was observed to be present) between 5% and 95%. Among them, 57 taxa are resolved to family level, one to order level, one to class level and one to phylum level. We can expect taxa on the family level or even coarser taxonomic levels to be less sensitive to environmental factors than on species level (Caradima et al., 2020). However, for a reasonable application of machine learning algorithms it is important to have a large enough sample size, therefore using smaller data sets with a better taxonomic resolution was not an option for this study. This is aligned with the Swiss assessment method for macroinvertebrates, which is based on family level (OFEV, 2019).

2.2. Environmental dataset

We selected nine environmental factors as explanatory variables (Table 1, see Supporting Information A 2.1 – A 2.3 for more details) based on expert knowledge and experience with statistical models previously applied to a similar macroinvertebrate dataset (Caradima et al., 2019, 2020). The data was derived from the Swiss Federal Office for the Environment (water quality monitoring data, hydrological data,

Table 1
Explanatory variables selected to predict the probability of occurrence of freshwater macroinvertebrates (see Supporting Information A 2.1 for more details on water temperature).

Environmental factor	Description
Temperature (°C)	Mean maximum morning summer stream temperature predicted from a linear mixed effect model based on catchment area, mean catchment elevation, and a random year effect
Flow velocity (m/s)	Mean annual stream flow velocity estimated from spatial data
Riparian agriculture (%)	Fraction of agricultural land use within a buffer distance of 10 m from the stream
Livestock unit density (CE/km^2)	Cattle equivalent (CE) units of livestock per square kilometer of catchment area
Insecticide application rate (-)	Sum of crop type fractions in the catchment weighted by the number of crop type specific insecticide treatments per year
Urban (%)	Proportion of urban and transport-related land use within the catchment
Forest (%)	Proportion of forest intersecting the river length in the catchment upstream from the site
Forest buffer (%)	Proportion of forest intersecting the river length within 150 m distance of the site
Width variability (-)	Expert evaluation of width variability of the stream channel ranging from 0 (channelized river) to 1 (natural width variability)

hydromorphological data, land use data), the Swiss Federal Office of Topography (topographical data), and the Swiss Federal Statistical Office (population statistics). We standardized each environmental factor by subtracting the mean and dividing by the standard deviation before applying the different models. We back-transformed them when analyzing the outputs of the models to facilitate the interpretation.

2.3. Model definitions and properties

We selected eight prototypical models based on their different properties and complexity as summarized in Table 2. By complexity we refer to the flexibility of the model structure imposing more or less constraints in the response shapes. Accordingly, we characterized the models by three properties: 1) enforcement of a smooth response shape, which seems desirable from a biological point of view, 2) the inclusion of potential interactions between environmental factors (e.g., the response to temperature could be different a low flow velocity than at high flow velocity), and 3) if a model predicts the occurrence of a single taxon or multiple taxa. Multi-taxa models may have an advantage for taxa with lower prevalence, if they can use information from other taxa (Caradima et al., 2019). The aim was to cover a wide range of model complexities. Therefore, we did not include interaction terms or more than quadratic transformations in the GLMs, because highly non-linear models are already covered by the ML approaches. Similarly, we did not run the multi-taxa models in a single taxon mode, because we are interested, if the predictive performance increases for multi-taxa models compared to single taxon models that are already included in the study.

The hGLM and chGLM (called UF0 and CF0 in Caradima et al., 2019) refer to hierarchical generalized linear models that are fitted jointly to multiple taxa, and where the regression coefficients of each taxon are constraint by an overarching community distribution. This prevents overfitting for taxa with unbalanced data (i.e., with low or high prevalence). The chGLM imposes in addition correlations among the community parameters that are inferred jointly with the regression coefficients.

The classification into statistical and ML models is somewhat arbitrary, because the transition is rather continuous. However, for easier comprehensibility, in the following we refer to the three GLMs and the GAM as statistical models and to the SVM, BCT, RF and ANN as ML models.

All models were interfaced from R using different packages and methods (see Table 2). The implementation of the hGLM and chGLM was based on Caradima et al. (2019). We took advantage of the caret R-package to automatically tune the hyperparameters of the GAM, SVM, BCT, and RF. The hyperparameters of the ANN were tuned manually (see Supporting Information A 3.3). The hyperparameters were tuned by three-fold cross-validation on the calibration dataset (see paragraph 2.4 and Supporting Information A 3.1). For each model, we used the same environmental factors listed in Table 1 as explanatory variables. We included a null model for comparison to analyze how much more the models learned compared to predicting only based on prevalence.

2.4. Model assessment

We evaluated the predictive performance of the models during three-fold cross-validation (CV) and during out-of-domain generalization (ODG). For CV, the data was randomly split into three subsets of equal size while taking care of avoiding data leakage, i.e., making sure that all observations from a site occur only in a single set (Nisbet et al., 2009). For ODG, we split the data into two subsets according to the temperature factor using the observations from the 80% coldest sites for calibration and the 20% warmest sites for prediction. We chose temperature, because it is the most important influence factor of aquatic macroinvertebrates (Ward and Stanford, 1982; Caradima et al., 2020) and of interest regarding climate change.

We assessed the predictive performance of the models by calculating the standardized deviance and, to allow for comparison with other

Table 2
Abbreviations, description, properties and implementation of the eight selected models plus the null model, sorted by increasing complexity. Note that we included a quadratic term for the factors temperature and flow velocity in the three different GLMs to account for taxa with mid-range preferences for temperature and flow velocity but we did not include any interaction terms.

Model	Description	Properties			Implementation	
		Smooth response shape	Interactions between environmental factors included	Single taxon or multi-taxa prediction	Package	Method
Null model	Probability of occurrence at all sites equals the prevalence	–	–	–	–	–
iGLM	Individual Generalized Linear Model (GLM)	yes	no	single	Caret	'glm'
hGLM	Hierarchical multi-taxa GLM	yes	no	multiple	Rstan	–
chGLM	Correlated hierarchical multi-taxa GLM	yes	no	multiple	Rstan	–
GAM	Generalized Additive Model	intermediate	no	single	Caret	'gamLoess'
SVM	Support Vector Machine	yes	yes	single	Caret	'svmRadial'
BCT	Boosted Classification Tree	no	yes	single	Caret	'ada'
RF	Random Forest	no	yes	single	Caret	'rf'
ANN	Artificial Neural Network (Multilayer Perceptron)	yes	yes	multiple	Keras	'keras_model_sequential'

studies, by measuring the area under the receiver operating characteristic curve (AUC) (Pearce and Ferrier, 2000). The AUC ranges from zero to one with higher values indicating a better performance. The standardized deviance is a statistical equivalent to the mean square of the residuals of a model with normally distributed errors (Hardin et al., 2007). It also is proportional to the binary cross-entropy used in the ML field. The standardized deviance is based on the likelihood for presence-absence data. We write $y_i = 1$ if a taxon is observed present in the i^{th} observation and $y_i = 0$ if it is observed absent. We use the upper case letter to describe the respective random variable Y_i. For a given taxon, model, its parameters $\boldsymbol{\theta}$, and the set of explanatory variables $\mathbf{x_i}$, the likelihood function is defined as:

$$P(y_i|\mathbf{x_i}, \boldsymbol{\theta}) = \begin{cases} P(Y_i = 1|\mathbf{x_i}, \boldsymbol{\theta}) \text{ if } y_i = 1, \\ 1 - P(Y_i = 1|\mathbf{x_i}, \boldsymbol{\theta}) \text{ if } y_i = 0. \end{cases}$$

The standardized deviance is calculated as follows:

$$d = \frac{-2}{n} \sum_{i=1}^{n} \log(P(y_i|\mathbf{x_i}, \boldsymbol{\theta}))$$

with n being the number of observations. By averaging over the number of observations, we can compare the performance of a model for taxa differing in their sample size. This metric ranges from zero to infinity, with a small value corresponding to a good performance. For further analysis and visualizing performance results, we chose to look only at the standardized deviance over other usual classification metrics to avoid defining an arbitrary classification threshold and to keep information on the predicted probability of occurrence (Lobo et al., 2007).

To quantify the degree of overfitting, we computed the likelihood ratio between calibration and prediction performance for each model (Hardin et al., 2007). For each taxon and each model, we calculate it as follows:

$$\lambda = e^{\frac{-d^{\text{cal}} - d^{\text{pred}}}{2}}$$

It corresponds to the ratio between the geometric mean of the likelihood during calibration and during prediction. It usually ranges from zero to one, with a lower value indicating a big gap in performance between the calibration and the prediction phases that we interpret as high degree of overfitting. If it is close to one, it indicates that the model performed similarly during calibration as during prediction. This metric could exceed one, if the predictive performance is better than the performance during calibration.

2.5. Taxon-specific response

To visualize the inferred response shapes between the predicted probability of occurrence of the taxa and the environmental factors for each model, we used two model agnostic methods (Molnar, 2019; Lucas, 2020).

In a first step, we plotted the Individual Conditional Expectation (ICE) (Goldstein et al., 2014). To this end, we randomly selected 100 observations from the monitoring data set, each representing a specific combination of environmental factors (we confirmed that the influence of the random sampling is minor by testing different seeds). For each of these combinations, we then predicted and plotted the probability of occurrence across the whole range of one environmental factor at a time, while the others remained fix, leading to 100 lines. On ICE plots, interactions between factors are reflected by non-parallel lines. Therefore, ICE show within which range of the environmental factor and to what extent the model learned interactions as well as non-linear responses.

The average model response to changing one variable is visualized by the Partial Dependence Plots (PDP) (Friedman, 2001), which can be approximated by averaging the ICE lines. The maximum difference of the PDP line on the y-axis across the whole range of the variable can be interpreted as a measure of sensitivity (as indicated by a black arrow in the figures). The PDP lines were also compared across all models to visualize differences in the learned responses.

If the environmental factors are independent from each other (e.g., for the GAM model), the PDP is exactly the averaged predicted responses, which makes it straightforward to interpret. As additional information, we also plotted the partial response when the other environmental factors are kept at their averaged value, as it is sometimes done in the SDM literature (e.g., Elith et al., 2008).

3. Results

3.1. Predictive performance during cross-validation

During cross-validation (CV), more complex models tend to predict slightly better than the statistical ones, both when assessed with the standardized deviance and the AUC. This can be observed in Table 3, with RF showing the best performance statistics for both metrics and in Fig. 1a from the green line showing that RF has the lowest median during prediction, followed by BCT, GAM and ANN. Fig. 1a also shows that all models have a better predictive performance than the null model based on the standardized deviance across all taxa.

We also see that more complex models, especially RF, demonstrate stronger overfitting, observable by the large difference in performance between calibration and prediction (distance between purple and orange boxes in Fig. 1a). This is also indicated by the median likelihood ratios (Table 4), where a lower value (e.g., 0.76 for RF) indicates a higher degree of overfitting compared to the statistical models (e.g., 0.99 for iGLM, hGLM and chGLM).

When looking at the predictive performance measured with the standardized deviance per taxon and sorted by prevalence (Fig. 2b), we see that for taxa with unbalanced data (i.e., low or high prevalence), most models are close to the null model. This is indicated by the colored dots that are close to the continuous black line, which represents the null model.

There is an improvement in predictive performance compared to the null model mainly for taxa with intermediate prevalence. However this improvement varies by taxon, with all models performing considerably better than the null model for some taxa, e.g., Gammaridae, and not for others, e.g., Psychodidae (Fig. 2b). Moreover, we observe only minor differences in predictive performance among models for the same taxon, as shown by the colored dots aligned vertically being close to each other in Fig. 2b. The figure also illustrates that the more complex models perform much better during calibration (Fig. 2a) than during prediction, especially RF represented by pink dots that are below 0.5 for all taxa, which again indicates strong overfitting.

Finally, we can visualize the geographic distribution of the predicted probability of occurrence of each model for each taxon during CV, where observations are colored in red if the taxon was observed absent and in blue if it was observed present (Fig. 3 shows predictions of the null model and RF for Gammaridae, see Supporting Information A 5 for the predictions of all models). Fig. 3 illustrates that RF has a better predictive performance than the null model, which shows the same probability of occurrence at all sites, as indicated by the size of the dots. However, when comparing RF with the other models, we observe only minor spatial differences in prediction, indicating again that the different models had a similar predictive performance during CV (Figure SI A 21).

3.2. Out-of-domain generalization

Averaged over all taxa, all models fail to generalize out-of-domain. This is indicated by the median of the distribution of the standardized deviance during prediction (black line in the orange boxes on Fig. 1b), being similar or even higher than the null model for every model. We can also see that overfitting of all models increased compared to CV, with RF again showing the strongest overfitting. This can also be seen in Table 4: all models have a lower median likelihood ratio (indicating

Table 3
Summary statistics of predictive performance measured with the standardized deviance (the lower the better) and the AUC (the higher the better) over all taxa for each model during cross-validation. Models are sorted from left to right by increasing level of complexity. Numbers in bold show the best performing model for each statistic, here RF for both metrics.

		iGLM	*hGLM*	*chGLM*	*GAM*	*SVM*	*BCT*	*RF*	*ANN*
Standardized deviance	min	0.38	0.39	0.38	0.39	0.41	0.38	**0.37**	0.40
(the lower the better)	median	0.81	0.81	0.82	0.80	0.86	0.79	**0.76**	0.81
	max	1.35	1.35	1.35	1.32	1.30	1.30	**1.28**	1.34
AUC	min	0.57	0.55	0.52	0.59	0.57	0.56	**0.61**	0.51
(the higher the better)	median	0.76	0.74	0.75	0.77	0.73	0.77	**0.78**	0.74
	max	0.91	0.91	0.91	**0.92**	0.91	**0.92**	**0.92**	0.90

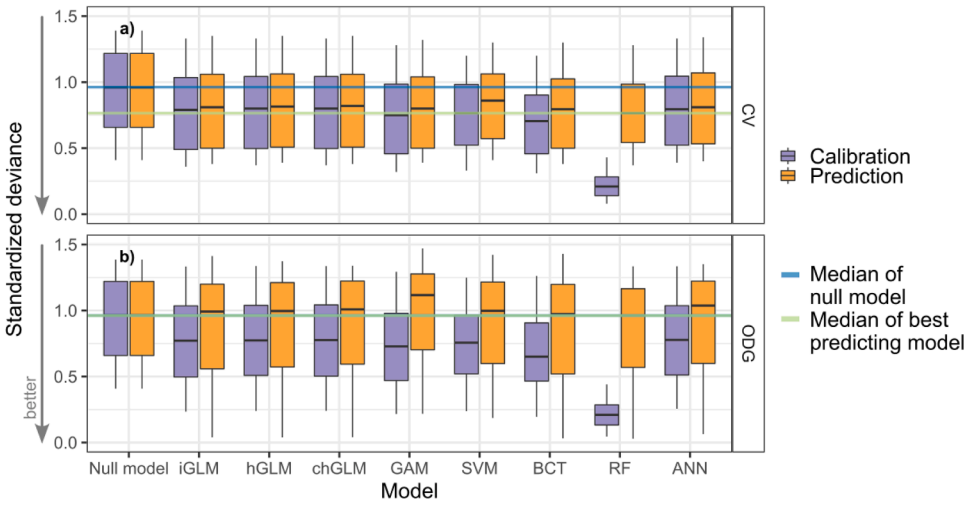


Fig. 1. Boxplots for the distribution of the standardized deviances (the lower the better) across the 60 taxa during calibration (purple) and prediction (orange), showing model performance. The top and the bottom of the boxes indicate the 75th and the 25th percentiles, respectively, of the distribution, while the horizontal black line in each box represents the median. Models on the x-axis are sorted by increasing level of complexity. The upper panel (a) shows cross-validation (CV) results, the lower (b) the out-of-domain generalization (ODG). The blue horizontal line shows the median standardized deviance of the null model and the green horizontal line shows the median standardized deviance of the best performing model during prediction, here RF during CV (a) and the null model during ODG (therefore, the blue and green line are overlapping) (b).

Table 4
Summary statistics for overfitting: the likelihood ratio over all taxa for each model during cross-validation and out-of-domain generalization. A lower value indicates a bigger difference between performance during calibration and prediction, interpreted as a higher degree of overfitting. Models are sorted from left to right by increasing level of complexity.

		iGLM	*hGLM*	*chGLM*	*GAM*	*SVM*	*BCT*	*RF*	*ANN*
Cross-validation	min	0.98	0.98	0.98	0.96	0.93	0.93	0.65	0.98
	median	0.99	0.99	0.99	0.97	0.96	0.96	0.76	0.99
	max	1	1	1	0.98	0.99	0.99	0.87	1
Generalization	min	0.61	0.6	0.59	0.11	0.52	0.57	0.57	0.62
	median	0.98	0.98	0.97	0.77	0.96	0.96	0.69	0.96
	max	1.29	1.29	1.29	1.25	1.24	1.26	1.07	1.3

higher overfitting) during ODG than during CV, with RF having the lowest value of 0.69. However, if we examine the predictive performance of the models for each taxon (Fig. 2d), we see that for most taxa with intermediate prevalence the models still perform better than the null model (as indicated by dots below the black line), while for most taxa with unbalanced prevalence the models perform even worse than the null model, in particular the models GAM and SVM, as indicated by the dots above the black line.

3.3. Comparison of learned response shapes

The Individual Conditional Expectation (ICE) and Partial Dependence Plots (PDP) illustrate the inferred responses of the taxa to the environmental factors of the different models. Figs. 4 and 5 show as an example the model response shapes of Gammaridae for temperature (see Supporting Information A 6 and B for the response of Gammaridae to all environmental factors and Supporting Information C for the response of all taxa to temperature).

We did not include interaction terms between environmental factors in the statistical models iGLM, hGLM, chGLM and GAM, which is indicated by the colored lines that do not cross each other in Fig. 4. In contrast, the SVM, BCT, RF and ANN account for interactions and learned different response patterns for the same environmental factor depending on the other factors, as indicated by the intersecting colored lines in Fig. 4. When generalizing on scarcer data, the models with constrained response shapes, i.e., iGLM, hGLM, chGLM and GAM, tend to an extreme prediction value, e.g., zero probability of occurrence for low temperature values. The more flexible models either allow for various responses for the same range of scarcer observations, e.g., SVM and ANN, or predict a constant value from the last observation, e.g., BCT and RF (Figs. 4 and 5). Overall, while the statistical models show unimodal response shapes, the other models show multiple local optima in the ICE (Fig. 4), and BCT and RF show large jumps in predicted probability of occurrence for small changes in temperature that are not plausible from an ecological point of view. While it would be possible that families consisting of species with distinct realized niches indeed exhibit multiple optima and the structure of tree-based models like RF and BCT allows non-smooth responses, the high degree of overfitting of these models indicates that these irregular shapes are rather artefacts of overfitting than ecologically plausible responses. Furthermore,

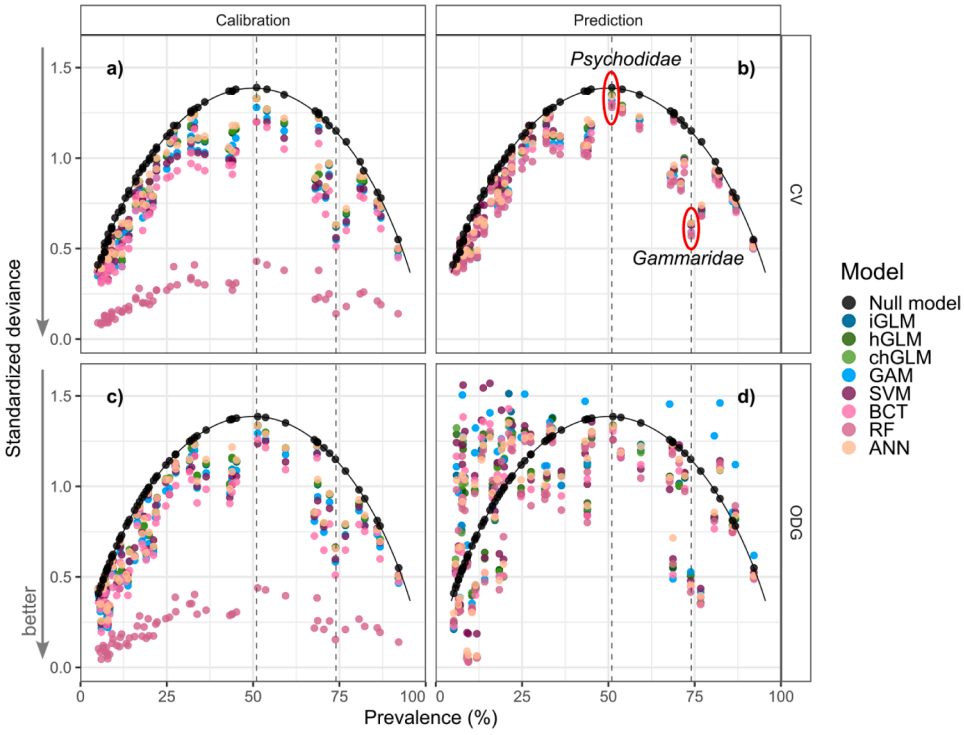


Fig. 2. Standardized deviance (the lower the better) of each model for each taxon, ordered on the x-axis according to their prevalence (only taxa with prevalence between 5% and 95% included), during calibration (left column) and prediction (right column), and during cross-validation (top row) and out-of-domain generalization (bottom row). The continuous black line indicates the null model performance. The vertical dashed lines show the position of Psychodidae (left) and Gammaridae (right) in each panel (see panel b for the labels).

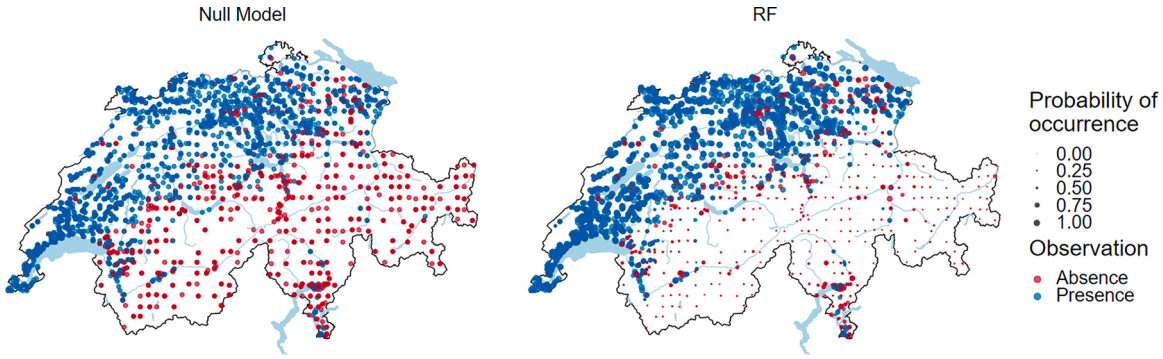


Fig. 3. Geographic distribution of the predicted probability of occurrence (indicated by the point size, see legend) of the null model (left map) and RF (right map) for Gammaridae during cross-validation. Observations are colored in red for an absence and in blue for a presence. Small red and large blue dots indicate a good agreement between model prediction and observation, which is mostly the case for the RF but not for the null model.

especially the ICE of the RF model show strong and irregular interactions for some sites that are not plausible.

A comparison of the partial dependence plots for temperature (PDPs, Figs. 4 and 5, see Figure SI A 24 for the other environmental factors) indicates that on average all models learned that the probability of occurrence of Gammaridae increases with temperature. Another common pattern among models is that the predicted probability of occurrence stabilizes above 17 °C. However, we can also observe that the complexity of the model and its way to generalize on scarcer data impacts the averaged responses. Thus, the statistical models present very similar smooth PDPs with a predicted probability of occurrence ranging from zero to one, while the more complex models predict a higher probability of occurrence for low temperature values, with BCT and RF predicting almost 0.5 probability of occurrence close to 0 °C and big jumps for higher temperature values, which is not plausible from an ecological point of view.

The difference between the minimum and the maximum values of the PDPs (vertical black arrows on the right of each panel in Fig. 4) indicates the average change in predicted probability of occurrence over the whole range of the explanatory variable. This is a measure of how sensitive the model is to a variable. Similarly, the range of the y-axis covered by each of the colored lines indicates the local sensitivity of the model at different points in the space of the other environmental predictors. Especially the RF, BCT and SVM exhibit a much smaller sensitivity to temperature compared to the GLMs, which cover almost the whole range from 0 to 1.

Finally, we observe that partial responses show different results depending on how they are calculated. Apparently, it matters if we first propagate the inputs through the model and average afterwards over the model outputs (as in the PDP, represented by thick black line in Fig. 4) or if we first average over the inputs and then propagate it through the model (as in the partial responses, thin black dashed line in Fig. 4).

4. Discussion

SDMs can inform management decisions by making predictions and by helping to increase our understanding of ecological systems (Srivastava et al., 2019; Urbina-Cardona et al., 2019). However, we need

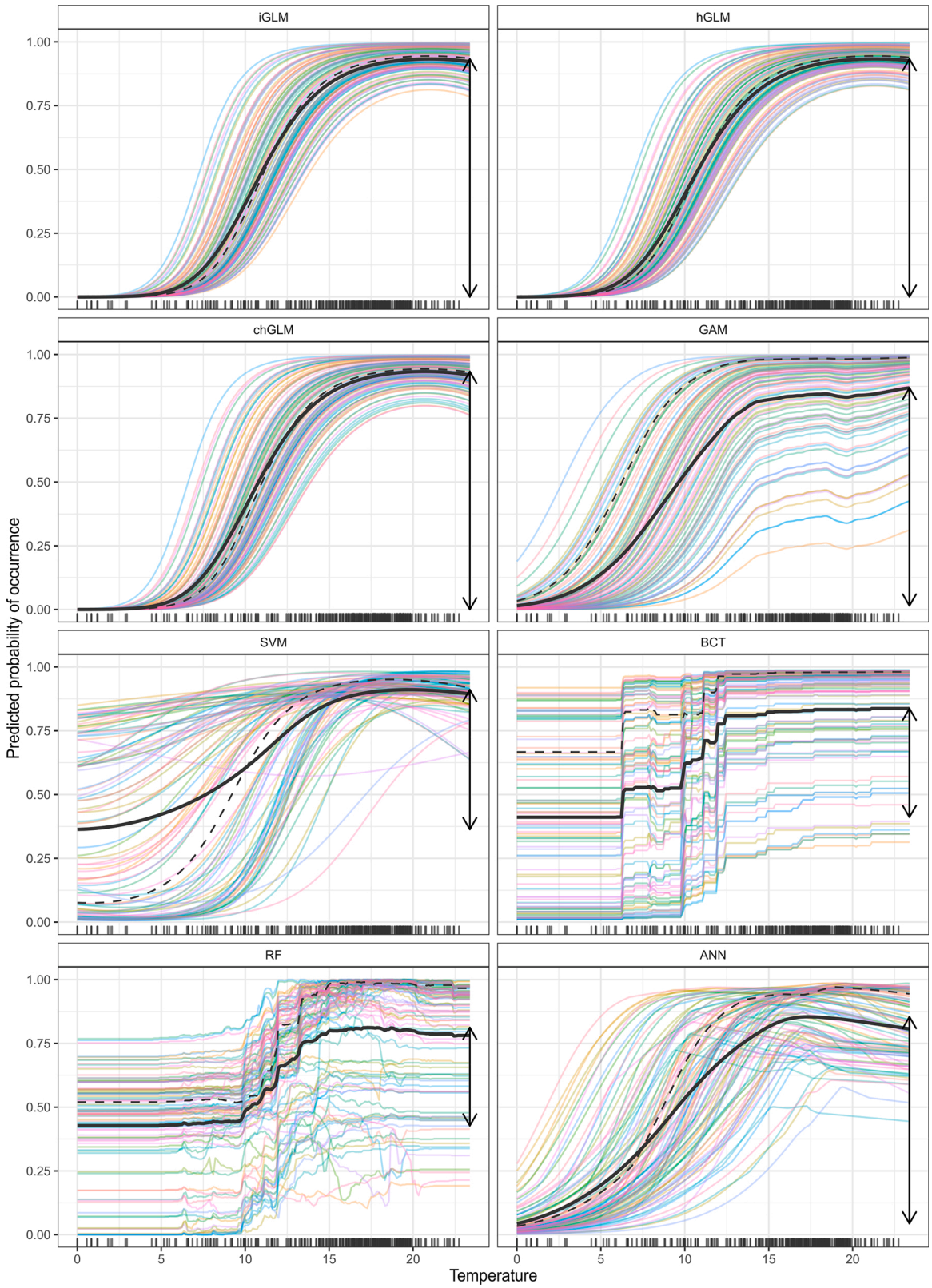


Fig. 4. Individual conditional expectation (ICE, colored lines), partial dependence plots (PDP, thick black line) and partial response when other factors are held at their mean (thin black dashed line) of inferred Gammaridae responses to temperature for the different models (panels) calibrated to the whole dataset and sorted from top to bottom by increasing level of complexity. Each colored line represents one of the 100 randomly selected observed combinations of environmental influence factors from the monitoring data. Intersecting lines indicate that the model learned interactions between environmental factors. The rug, i.e., small vertical lines on the x-axis, represents the distribution of temperature values in the 100 randomly selected observations. It shows how the different models generalize when the observations are scarcer, e.g., for temperatures values below 10 °C or above 21 °C. The black vertical arrow on the right of each panel represents the difference between the minimum and the maximum values of the PDP, indicating the sensitivity of the model to the variable.

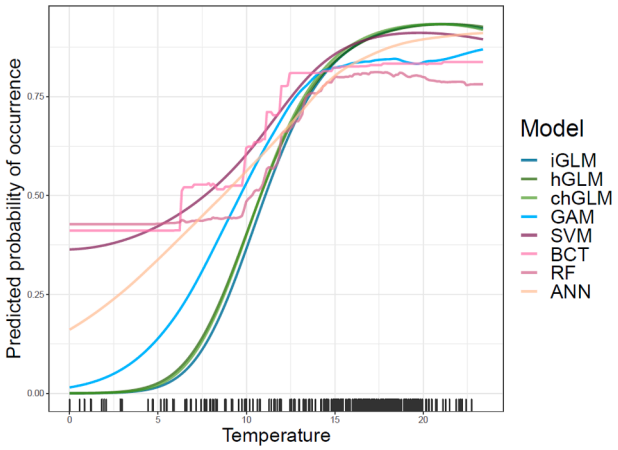


Fig. 5. Comparison of Partial Dependence Plots of inferred Gammaridae responses to temperature among models (see legend for the color-coding). The rug, i.e. small vertical lines on the x-axis, represents the distribution of temperature values in the 100 randomly selected observations used to calculate the PDPs.

confidence that the model learned the right relationship between causes and effects (Schuwirth et al., 2019). To increase confidence, it is important to 1) select the explanatory variables based on the knowledge about causal effects in the system (Arif and MacNeil, 2022), and to 2) assess the plausibility of the response shapes inferred by the model.

4.1. Minor effects of increasing complexity on predictive performance

In this study, we compared models with different complexity to understand how complexity affects predictive performance and inferred responses. We were expecting ML models, such as RF, to have a better predictive performance during cross-validation (CV) than simple statistical models, because their structure allows them to learn more flexible response shapes (Li and Wang, 2013). By using increasingly complex models with different properties, we expected to learn, which additional features could be included in the simpler models to gain predictive performance while keeping an easily interpretable structure. However, for this data set and by looking at the median over all taxa, the more complex models showed only a minor improvement in predictive performance compared to the statistical models (chGLM and RF have only a difference of 0.03 in median AUC and 0.06 in median standardized deviance in prediction during CV). This statement is supported by the comparison of the geographic distribution of the predicted probability of occurrence for each taxon (Supporting Information A 5), where we observe only minor differences among models. Only for some specific taxa with intermediate prevalence the ML models performed substantially better than the simpler statistical models. This indicates that a larger number of data points could improve the inference of more complex responses, especially for the rarer taxa.

During out-of-domain generalization (ODG), the statistical models were expected to have a better predictive performance than the ML models, because they are more constrained. However, by looking at the median standardized deviance over all taxa, all models performed similarly or worse than the null model. This indicates that we have to be careful, especially for taxa with unbalanced prevalence, if we want to use any of the tested SDMs to predict effects of environmental changes beyond the range that is covered in the calibration data, for example the effect of global warming for sites that are already at the upper end of the temperature range. If it is required to make predictions out of the calibration domain, and knowledge about taxa responses under these circumstances is available from other sources, a model that can include prior knowledge about the preferences of the taxa could be useful (e.g., Vermeiren et al., 2020b). For some taxa with intermediate prevalence, the out-of-domain generalization worked better than expected and even better than under cross-validation (for example Gammaridae). This seems to be related to the prevalence in the ODG training (calibration) and testing (prediction) data set (see Supporting Information A 3.2). For example, Gammaridae had a higher prevalence in the ODG testing data set than in the training. They occurred at most of the warmest sites. This was easier to predict by the model, because the probability of occurrence at the upper end of the training data was already high, so that extrapolation lead to a very high performance. For other taxa, such as Nemouridae, the optimum temperature was around the temperature threshold that was used to divide training and testing data. For this taxon, the models were not able to predict the decreasing probability of occurrence at the highest temperatures and therefore performed worse during ODG.

4.2. Effect of overfitting on interpretability

While the ML models only partially predicted better than the statistical models in this study, the main impact of their complexity is observed on overfitting and consequently on the learned response shapes and their interpretability. The difference between performance during calibration and prediction was much higher for the more complex models (especially RF with an average likelihood ratio of 0.76) than for the simpler GLM models (average likelihood ratio of 0.99), which indicates that the complex models fit closer to the data. When taking the RF alone, the large degree of overfitting makes it difficult to judge, which characteristics of the response shapes capture real patterns and which are just describing the noise in the data. For this, a comparison with models that are less overfitting but have a similar predictive performance are helpful.

The learned response shapes that are visualized with the ICE and PDP plots indicate that more flexibility allows the more complex models to learn interactions and complex relationships, but also sometimes lead to multiple optima and abrupt jumps. Ecological theory supports that response shapes should be unimodal (with the exception of taxa consisting of multiple species with distinct responses) and smooth (Oksanen, 1997; Austin, 2007; Holt, 2009), which questions the plausibility of the response shapes inferred by the more complex models and their suitability for learning about the true system behavior. It is known that the properties of some models, in particular the tree-based structure of RF and BCT, result in non-smooth response shapes. As even more critical we judge the sometimes strong interactions. In order to support environmental decisions, we often want to make site-specific predictions. While we can expect some interactions in the response to some variables, the ICE for the RF model shows that the influence of a variable can be inverted for a few sites, which seems implausible. Since these models also show stronger overfitting and similar predictive performance than the statistical models (with lower degree of overfitting), it is justified to interpret the implausible responses shapes and interactions as results of fitting the noise in the data.

Finally, we observe that also the choice of visualization can impact the interpretation. For the same taxon, we can get different response shapes for the PDP compared to the partial response calculated by setting the other factors to their mean (see for instance the difference between the straight and dotted black line in the GAM and other more complex models in Fig 4.). The computation of the PDP is more expensive, but it considers a broader range of input data and is therefore more informative.

4.3. Choosing the right level of complexity

In summary, our results indicate that identifying models with a big difference in performance during calibration and prediction, i.e., a high degree of overfitting, is a good indicator for implausible complex response shapes. In this study we found that more flexible response shapes are not necessary to describe the patterns in the data or that the

underlying complex relationships are masked by noise. Overall, it is acknowledged that it is challenging to find the optimal level of complexity to reach the best predictive performance (Elith and Franklin, 2013; Merow et al., 2014). This is illustrated in Fig. 6, where the change in performance during calibration and prediction is plotted against model complexity. Because plausible response shapes are important to interpret the results of a model with confidence, we would prefer to select a model that is slightly underfitting (e.g., point A in Fig. 6) than a model having a bigger difference in performance between calibration and prediction and consequently showing overcomplicated response shapes, even if it is predicting slightly better (e.g., point B or point C in Fig. 6). For instance, in our study, RF presents a large difference in performance during calibration and prediction and implausible response shapes, while predicting only slightly better than simpler statistical models, indicating that the RF is closer to point C than to point A.

4.4. Limitations and outlook

In this study, the overfitting problem is most likely caused by the noise in the dataset. It highlights that the limiting factor for further improving predictive performance is not model complexity but the information content of the response and explanatory variables. Increasing the amount of data, especially for taxa with low prevalence, and improving the precision of the environmental factors used as explanatory variables could help increasing the predictive performance of all models. For some influence factors, e.g., water quality, we had to rely on proxies based on land use. More directly linked variables, e.g., measurements of chemical water quality, could improve the predictive performance. Including other spatial factors might also lead to an improvement in predictive performance, e.g., by accounting for dispersal limitation. Similarly, improving the taxonomic resolution in the data, e.g., from family to genus or species level, could lead to an improvement in predictive performance because the models could learn stronger responses, especially for taxa with a high prevalence (Caradima et al., 2019, 2020; Vermeiren et al., 2020a, 2020b). However, a better taxonomic resolution could also increase the noise in the data in some cases, due to misidentification, and would lead to a decrease in prevalence.

For future studies with similar data, we recommend to put more effort in preventing ML models to overfit by applying stronger regularization. For the scope of this paper, we intended to use standard approaches implemented in software packages that are currently used in the field of SDM, such as the R-package "caret". The "caret" package uses internal tuning or basic grid search (Kuhn, 2022), which did not lead to a significant improvement in prediction by hyper-parameter tuning. A better regularization to find an optimum level of complexity so that overfitting is minimized while still reaching a high predictive performance (see Fig. 6) would increase the confidence in interpreting ML model results. It may be promising to investigate more complex tree-based models that allow to regularize the interactions strength, for example XGBoost (Chen and Guestrin, 2016).

Other ways to deepen the analysis of taxa responses to environmental factors would be the use of time series data with a temporal resolution that fits the temporal dynamics and the integration of information from controlled experiments, ideally from the field. The latter could be achieved by a monitoring design that is targeted to success control of large scale management programs.

Future research might also explore the use of synthetic data with different noise levels to discern the sensitivity of the different modeling approaches to noise (Austin et al., 2006; Meynard and Quinn, 2007; Zurell et al., 2010). This could help to draw generalizable conclusions about the optimal level of model complexity depending on the noise in the data in a controlled setting.

As a general conclusion, to find the optimal level of model complexity in the context of SDMs, we recommend the following steps:

1) Apply several models that differ in complexity.
2) Compare their predictive performance and degree of overfitting, as indicated by the difference in performance during calibration and prediction.
3) Compare the inferred response shapes with model agnostic tools, such as PDP and ICE plots.
4) Evaluate, if the response shapes are ecologically plausible by putting them in the perspective of the results in predictive performance and degree of overfitting.

This comparison of predictive performance, overfitting, and learned response shapes allows us to understand, if a gain in predictive performance is due to real complexity in the relationships between the explanatory and response variables, or rather due to fitting to the noise in the data. Such a comparison helps us to make a well-grounded choice on the level of model complexity.

5. Conclusions

Despite the increasing use of machine learning algorithms in species distribution modeling, some work still needs to be done to determine, which model is most appropriate for the data available and the purpose of a study (Tredennick et al., 2021). Our study shows that the comparison of several modeling approaches with different levels of complexity and their resulting response shapes is a useful approach to gain confidence in model predictions. We found that, for this specific dataset, the improvement in predictive performance between the statistical and the machine learning models is too small to accept overcomplicated response shapes. The degree of overfitting was a good indicator that the

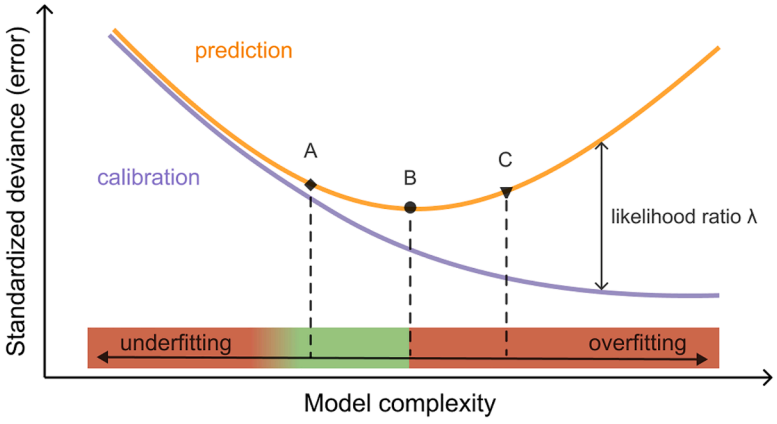


Fig. 6. Conceptual illustration of the dependence of predictive performance (orange line) and performance during calibration (purple line) based on the standardized deviance (y-axis) with a lower value indicating a better performance, on model complexity, represented conceptually on the x-axis. Overfitting is indicated by the difference between the two lines, the likelihood ratio λ (vertical arrow between orange and purple lines). The point B (black point) shows the optimal complexity regarding predictive performance. The point A (diamond) indicates a compromise between high predictive performance and low overfitting that could be expected to lead to more plausible response shapes. The point C (triangle) shows that even an intermediate predictive performance between point A and B can be related to a higher degree of overfitting, leading to unplausible response shapes.

response shapes should be interpreted carefully. The individual conditional expectation and partial dependence plots showed that more complex models fitted the noise in the data by allowing complex interactions, abrupt jumps and multiple optima in the response shapes for only a minor improvement in predictive performance, compared to simpler models showing less overfitting and smoother response shapes. Moreover, on average all models learned similar patterns in the response shapes, especially in the range of the explanatory variables that was well covered in the data, implying that statistical models already captured useful information in the data. Therefore, we should be careful when using data-driven approaches to learn about the system in an ecological sense. A critical review of inferred responses in light of ecological knowledge about the taxa is always recommended before using the results to inform decision making. In particular, we recommend to be cautious for predictions regarding future environmental changes that go beyond the range of environmental conditions observed so far. We conclude that a comparative analysis of models with differing complexity can guide model selection and increase the confidence in the interpretation of results.

Credit author statement

Author contributions: NS initiated the study and acquired funding. ECR, NS and AS designed the study. ECR and JW implemented the models and carried out the simulations and visualization of the results. All authors contributed to the analysis of the results. ECR wrote the first draft of the paper and all authors contributed to writing and revisions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and code are shared here: https://github.com/emmachollet/ComparSDMsQuantifOverfitSuppInterpr_DataPackage.

Acknowledgement

We thank the Swiss National Science Foundation (SNSF) for funding this study (grant 310030_192503). We thank the Swiss Federal Office for the Environment (FOEN), especially Yael Schindler Wildhaber, and the info fauna CSCF & karch, especially Maxime Chèvre, for access to the data and support. We acknowledge Rosi Siber for data preparation, Raoul Schaffner and Stuart Dennis for IT support, Janneke Hille Ris Lambers, Helen Moor and Catalina Chaparro Pedraza for helpful comments on an earlier version of the manuscript, and Peter Reichert, Marco Baity-Jesi, Emanuele Francazi, Marvin Höge, Florian Altermatt and Roman Alther for stimulating discussions throughout the development of this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2023.110353.

Bibliography

Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5, eaat4858. https://doi.org/10.1126/sciadv.aat4858.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. (Amst.) 22, 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Arif, S., MacNeil, M.A., 2022. Predictive models aren't for causal inference. Ecol. Lett. 25, 1741–1745. https://doi.org/10.1111/ele.14033.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Modell. 200, 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D., Luoto, M., 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. Ecol. Modell. Predict. Species Distribut. 199, 197–216. https://doi.org/10.1016/j.ecolmodel.2006.05.023.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423. https://doi.org/10.1111/j.1365-2664.2006.01136.x.

Beery, S., Cole, E., Parker, J., Perona, P., Winner, K., 2021. Species distribution modeling for machine learning practitioners: a review. In: ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS). Presented at the COMPASS '21: ACM SIGCAS Conference on Computing and Sustainable Societies. ACM, Virtual Event Australia, pp. 329–348. https://doi.org/10.1145/3460112.3471966.

Caradima, B., Reichert, P., Schuwirth, N., 2020. Effects of site selection and taxonomic resolution on the inference of stream invertebrate responses to environmental conditions. Freshwater Sci. 39, 415–432. https://doi.org/10.1086/709024.

Caradima, B., Schuwirth, N., Reichert, P., 2019. From individual to joint species distribution models: a comparison of model complexity and predictive performance. J. Biogeogr. 46, 2260–2274. https://doi.org/10.1111/jbi.13668.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Elith, J., Franklin, J., 2013. Species distribution modeling. In: Levin, S.A. (Ed.), Encyclopedia of Biodiversity, 2nd Edition. Academic Press, Waltham, pp. 692–705. https://doi.org/10.1016/B978-0-12-384719-5.00318-X.

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annal. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

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. Modell. Softw. 47, 1–6. https://doi.org/10.1016/j.envsoft.2013.04.005.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2014. Peeking Inside the black box: visualizing statistical learning with plots of individual conditional expectation. [stat].

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Modell. 157, 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1.

Hardin, James W., Hardin, James William, Hilbe, J.M., Hilbe, J., 2007. Generalized Linear Models and Extensions, 2nd Edition. Stata Press.

Holt, R.D., 2009. Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. Proc. Natl. Acad. Sci. 106, 19659–19665. https://doi.org/10.1073/pnas.0905137106.

IPBES, 2019. Global Assessment Report On Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Zenodo. https://doi.org/10.5281/zenodo.6417333.

Kuhn 2022, M.. caret: Classification and Regression Training. R package version 6.0-93. https://CRAN.R-project.org/package=caret.

Li, X., Wang, Y., 2013. Applying various algorithms for species distribution modelling. Integr. Zool. 8, 124–135. https://doi.org/10.1111/1749-4877.12000.

Linke, S., Norris, R.H., Pressey, R.L., 2008. Irreplaceability of river networks: towards catchment-based conservation planning. J. Appl. Ecol. 45, 1486–1495. https://doi.org/10.1111/j.1365-2664.2008.01520.x.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2007. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

Lucas, T.C.D., 2020. A translucent box: interpretable machine learning in ecology. Ecol. Monogr. 90, e01422. https://doi.org/10.1002/ecm.1422.

Merow, C., Smith, M.J., Edwards Jr, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281. https://doi.org/10.1111/ecog.00845.

Meynard, C.N., Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J. Biogeogr. 34, 1455–1469. https://doi.org/10.1111/j.1365-2699.2007.01720.x.

Møller, A., Jennions, M.D., 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132, 492–500. https://doi.org/10.1007/s00442-002-0952-2.

Molnar, C., 2019. Interpretable machine learning, a guide for making black box models explainable [WWW Document]. URL https://christophm.github.io/interpretable-ml-book/(accessed 3.4.21).

Nisbet, R., Elder, J., Miner, G., 2009. Chapter 20 - top 10 data mining mistakes. In: Nisbet, R., Elder, J., Miner, G. (Eds.), Handbook of Statistical Analysis and Data Mining Applications. Academic Press, Boston, pp. 733–754. https://doi.org/10.1016/B978-0-12-374765-5.00020-6.

Oksanen, J., 1997. Why the beta-function cannot be used to estimate skewness of species responses. J. Veget. Sci. 8, 147–152. https://doi.org/10.2307/3237252.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 225–245. https://doi.org/10.1016/S0304-3800(00)00322-7.

Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P., Lees, D.C., 2006. Model-based uncertainty in species range prediction. J. Biogeogr. 33, 1704–1711. https://doi.org/10.1111/j.1365-2699.2006.01460.x.

Rahman, Md.S., Pientong, C., Zafar, S., Ekalaksananan, T., Paul, R.E., Haque, U., Rocklöv, J., Overgaard, H.J., 2021. Mapping the spatial distribution of the dengue vector Aedes aegypti and predicting its abundance in northeastern Thailand using machine-learning approach. One Health 13, 100358. https://doi.org/10.1016/j.onehlt.2021.100358.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33, 1689–1703. https://doi.org/10.1111/j.1365-2699.2006.01466.x.

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44, 199–205. https://doi.org/10.1111/ecog.05360.

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S.D., Martínez-López, J., Vermeiren, P., 2019. How to make ecological models useful for environmental management. Ecol. Modell. 411, 108784 https://doi.org/10.1016/j.ecolmodel.2019.108784.

Srivastava, V., Lafond, V., Griess, V.C., 2019. Species distribution models (SDM): applications, benefits and challenges in invasive species management. CABI Rev. 2019, 1–13. https://doi.org/10.1079/PAVSNNR201914020.

Stupariu, M.-S., Cushman, S.A., Pleşoianu, A.-I., Pătru-Stupariu, I., Fürst, C., 2021. Machine learning in landscape ecological analysis: a review of recent approaches. Landscape Ecol. https://doi.org/10.1007/s10980-021-01366-9.

Timoner, P., Fasel, M., Ashraf Vaghefi, S.S., Marle, P., Castella, E., Moser, F., Lehmann, A., 2021. Impacts of climate change on aquatic insects in temperate alpine regions: complementary modeling approaches applied to Swiss rivers. Glob. Chang. Biol. 27, 3565–3581. https://doi.org/10.1111/gcb.15637.

Tredennick, A.T., Hooker, G., Ellner, S.P., Adler, P.B., 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102, e03336. https://doi.org/10.1002/ecy.3336.

Tuanmu, M.-N., Viña, A., Roloff, G.J., Liu, W., Ouyang, Z., Zhang, H., Liu, J., 2011. Temporal transferability of wildlife habitat models: implications for habitat monitoring. J. Biogeogr. 38, 1510–1523. https://doi.org/10.1111/j.1365-2699.2011.02479.x.

Urbina-Cardona, N., Blair, M.E., Londoño, M.C., Loyola, R., Velásquez-Tibatá, J., Morales-Devia, H., 2019. Species distribution modeling in Latin America: a 25-year retrospective review. Trop. Conserv. Sci. 12, 1940082919854058 https://doi.org/10.1177/1940082919854058.

Vermeiren, P., Reichert, P., Graf, W., Leitner, P., Schmidt-Kloiber, A., Schuwirth, N., 2021. Schuwirth, N. (2021). Confronting existing knowledge on ecological preferences of stream macroinvertebrates with independent biomonitoring data using a Bayesian multi-species distribution model. Freshwater Science 40 (1). https://doi.org/10.1086/713175, 202-220.

Vermeiren, P., Reichert, P., Schuwirth, N., 2020b. Integrating uncertain prior knowledge regarding ecological preferences into multi-species distribution models: effects of model complexity on predictive performance. Ecol. Modell. 420, 108956 https://doi.org/10.1016/j.ecolmodel.2020.108956.

Visser, H., Evers, N., Bontsema, A., Rost, J., de Niet, A., Vethman, P., Mylius, S., van der Linden, A., van den Roovaart, J., van Gaalen, F., Knoben, R., de Lange, H.J., 2022. What drives the ecological quality of surface waters? A review of 11 predictive modeling tools. Water Res. 208, 117851 https://doi.org/10.1016/j.watres.2021.117851.

Ward, J.V., Stanford, J.A., 1982. Thermal responses in the evolutionary ecology of aquatic insects. Annu. Rev. Entomol. 27, 97–117. https://doi.org/10.1146/annurev.en.27.010182.000525.

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Method. Ecol. Evol. 3, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Werkowska, W., Márquez, A.L., Real, R., Acevedo, P., 2017. A practical overview of transferability in species distribution modeling. Environ. Rev. 25, 127–133. https://doi.org/10.1139/er-2016-0045.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., Grimm, V., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119, 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x.

Zurell, D., Elith, J., Schröder, B., 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. Divers. Distribut. 18, 628–634. https://doi.org/10.1111/j.1472-4642.2012.00887.x.

OFEV (ed.), 2019. Méthodes d'analyse et d'appréciation des cours d'eau (IBCH_2019). Macrozoobenthos – niveau R. 1ère édition actualisée 2019, 1re édition 2010. Office fédéral de l'environnement, Berne, L'environnement pratique no 1026: 58 p. URL: www.bafu.admin.ch/uv-1026-f (accessed 02.18.2021).