Integrating uncertain prior knowledge regarding ecological preferences into multi-species distribution models: Effects of model complexity on predictive performance

Peter Vermeiren⁎, Peter Reichert, Nele Schuwirth
Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

ARTICLE INFO

Keywords:
Bayesian inference
Ecological niches
Hierarchical modelling
Multiple stressors
Macroinvertebrates

ABSTRACT

Species distribution models (SDMs) are often criticised for lacking explicit linkage to ecological concepts. We aim to improve the ecological basis of SDMs by integrating prior knowledge about ecological preferences of organisms. Additionally, we aim to support a systematic, data-driven review of such prior knowledge by confronting it with independent monitoring data using Bayesian inference. We developed a series of multi-species distribution models (MSDMs) with increasing complexity to predict the probability of occurrence of taxa at sampling sites based on habitats suitability functions that are parameterized with prior ecological knowledge. We subsequently assessed the models' predictive performance with 3-fold cross-validation. So far, if ecological preferences or functional traits have been used in SDMs, they were mainly used as fixed inputs without considering their uncertainty. We take the additional step of considering uncertainty about preference parameters by including them as uncertain prior information that is subsequently updated with Bayesian inference. We apply the series of models in a case study on macroinvertebrates in Swiss streams. We analyse differences in the quality of fit, changes in predictive performance, and the potential to learn about the parameters from the data. We consider ecological preferences for natural and human modified environmental factors including temperature, flow velocity, organic matter concentration, insecticide pollution, and substratum. Results indicate that updating prior knowledge on ecological preferences with Bayesian inference, rather than using it as fixed input, improves model fit and predictive performance. For example, the predictive performance measured by the deviance for validation data improves by 17% and the explanatory power increases 3.8 times from a model that treats ecological preferences as fixed scores to a model that treats them as uncertain parameters. The spatial distribution of many taxa, including rare taxa with frequencies of occurrence down to about 5%, which are difficult to model with SDMs that do not consider prior information, can be captured by the new models. Integrating prior knowledge as uncertain parameters in a Bayesian framework establishes ecological interpretable links between taxa and their environment and supports a systematic revision and complementation of databases on ecological preferences, even in case of poor or missing prior knowledge. Model outputs need to be carefully interpreted by modellers and experts on ecological preferences. Increased exchange between these research fields will benefit further integration of ecological preferences into SDMs.

1. Introduction

Species Distribution Models (SDMs) are valuable tools in ecology and environmental management as they support our understanding of how natural and human factors affect species distribution patterns (Guisan and Zimmermann, 2000; Elith and Leathwick, 2009). A key strength of SDMs is their ability to represent distribution patterns at large, management relevant, scales (Guisan and Zimmermann, 2000). However, SDMs have been criticised for lacking explicit linkage to ecological concepts (Elith and Leathwick, 2009). Consequently, outputs of SDMs often do not increase our ecological knowledge and have only restricted management applicability (Guisan et al., 2013). To strengthen the link between the empirical and theoretical knowledge to be included in SDMs, these models need to consider cause-effect linkages between influence factors and species distributions, separate effects of multiple stressors, and produce predictions and knowledge that is applicable across regions and taxonomic compositions (Guisan et al., 2013).
Several databases synthesizing ecological knowledge, such as information on functional traits (sensu McGill et al., 2006) and ecological preferences (i.e. habitat requirements, ecological niches), are available for a wide range of organisms (e.g. Jones et al., 2009, Tachet et al., 2010, Kattge et al., 2011, Schmidt-Kloiber and Hering, 2015). Integrating such knowledge into SDMs can increase ecological interpretation of modelled patterns and the value of SDMs for management and conservation (Evans et al., 2015). Moreover, SDMs integrating ecological knowledge of taxa in addition to pure taxonomical descriptions potentially have a wider applicability (Loegel et al., 2013) and can assist in separation of effects from multiple influence factors (Mondy et al., 2012; Szöcs et al., 2014). Often, ecological databases pool knowledge from a variety of sources including controlled experiments and field observations through a process of literature review and expert validation (Schmidt-Kloiber and Hering, 2015; Serra et al., 2016). This compilation of existing ecological knowledge in databases facilitates its uptake and integration in SDMs. Especially knowledge on ecological preferences related to habitat requirements or sensitivity to natural and human influence factors is well suited for integration into SDMs as direct links with environmental influence factors can be made.

Currently, SDMs integrating ecological knowledge regarding functional traits or ecological preferences (e.g. RLQ and fourth corner ordinations, Dray et al., 2014, Generalized Joint Attribute Modelling, GJAM, Clark, 2016, Hierarchical Modelling of Species Communities, HMSC, Ovaskainen et al., 2017, or Hierarchical multi-SDMs and mixing models, Pollock et al., 2012; Jamil et al., 2013), have mostly treated this knowledge as known inputs (but see Bennetse et al., 2016 who modify this knowledge based on field data and expert opinions and then use it as model input). However, both functional traits and ecological preferences are uncertain and may vary for example between local populations (Vermeiren et al., 2015) or across life stages. Since functional traits can be observed at the level of individual organisms, it does not make sense to infer them from species occurrences. Ecological preferences, however, cannot be deduced from observations at the individual level, and there are several reasons why ecological preferences are considerably more uncertain than functional traits, and thus might benefit from being treated as uncertain parameters in SDMs.

First, selection processes at different spatial scales influence species distributions through environmental filtering (Poff, 1997). Consequently, species with ecological preferences that do not match the local conditions at a specific site will usually not occur at this site. However, additional factors influence species distribution patterns such as historical events, dispersal, biotic interactions and random processes that are not explicitly captured in ecological preference information (Leibold et al., 2004; Cadotte and Tucker, 2017). Consequently, the successful integration of ecological preferences into SDMs to represent and predict species distribution patterns needs to consider such additional processes. This has previously been achieved, for instance, via the use of mixed models or by adjustment for overprediction by a posteriori assembly rules (Guisan and Rahbek, 2011; D’Amen et al., 2015).

Secondly, the process of synthesising ecological preference information from various sources is not always traceable. Consequently, the exact interpretation of ecological preferences presented in databases, e.g. as an absolute limitation or rather as a looser restriction, can be difficult. Additionally, the actual boundaries between classes and even the link to a specific influence factor (e.g. minimum temperature vs. average temperature) can be vague.

A final challenge relates to data limitations. Direct knowledge on ecological preferences of species through experimentation and observation can be hard to collect due to resource limitations. However, mining existing literature and species records can reveal a large amount of information, which can be further augmented by expert opinions and traditional knowledge, albeit knowledge gaps persist for certain groups of organisms (Costello et al., 2010). Additionally, for organisms such as macroinvertebrates, the taxonomic resolution of monitoring data can vary among taxa (Damanik-Ambirita et al., 2018) and there can be a mismatch between the taxonomic resolution of the observational data and the preference information. Often, taxonomic or phylogenetic relations are used to assign functional traits and ecological preferences for species with missing data (e.g. Poff, 1997; Guérand et al., 2013; Penone et al., 2014). However, ecological preferences in particular can vary even within genera (Serra et al., 2016). Hence, while databases on ecological preferences contain valuable information, a straightforward method to support the validation and complementation of preference knowledge with independent data will be useful.

We aim to improve the ecological interpretability of SDMs by taking advantage of both existing prior knowledge regarding organisms’ ecological preferences, available in databases, and independently acquired monitoring data. Moreover, to contribute to the synthesis of ecological knowledge, we aim to support data-driven learning about ecological preferences by confronting existing prior knowledge from databases with independent monitoring data (that was not used for developing the preference databases). Therefore, we conduct a stepwise development of several multi-species distribution models (MSDMs), which explicitly utilize knowledge about ecological preferences to derive habitat suitability functions, and subsequently use these habitat suitability functions to predict the probability of occurrence of individual taxa that make up the community at any given site. We further refer to them as Habitat Suitability based Multi-Species Distribution Models (HS-MSDMs).

We start with a purely knowledge driven model based only on the prior information about ecological preferences from the databases and the related environmental influence factors as inputs. Subsequently, we estimate the relative importance of different environmental influence factors by assigning a weighting parameter to the habitat suitability regarding each influence factor and estimate its value using Bayesian inference. Finally, we treat the information about ecological preferences of taxa as uncertain prior knowledge about the parameters and allowing for a data-driven learning regarding ecological preferences. Specifically, we use Bayesian inference to update prior knowledge about all model parameters (now explicitly including parameters for ecological preferences of individual taxa) based on independent monitoring data (that has not been used earlier to derive the information about ecological preferences). To the best of our knowledge, our study is the first to integrate ecological preference information from databases as uncertain prior knowledge about the parameters to be updated using Bayesian inference in SDMs. Bayesian inference has been used successfully in other ecological research to build upon and revise current ecological knowledge (McCarthy, 2007; Hobbs and Hooten, 2015).

Species distribution models are well suited to describe taxon responses to environmental influence factors. However, it cannot be a priori known, if all important environmental influence factors are included in the model. Furthermore, additional factors influence community assembly, such as dispersal and biotic interactions (HilleRisLambers et al., 2012). The addition of a taxon-specific parameter that modifies the probability of occurrence of an individual taxon at all sites can to some degree account for dispersal limitation of rare taxa. In addition, we implement a site effect, which can modify the probability of occurrence of all taxa in the community jointly at each site. Analyzing spatial patterns of the site effect can facilitate the identification of additional factors or processes that can potentially improve the model.

By comparing the different versions of the HS-MSDMs, we address the following research questions: 1) What does the increased model complexity in each step add to the ecological knowledge that can be
gained from the HS-MSDM? 2) To which degree does increasing model complexity affect model fit and predictive performance? 3) How do spatial effects change model fit and predictive performance and do they support the identification of missing processes? We use freshwater macroinvertebrates to illustrate and test our model concept. Macroinvertebrates often display strong niche associations, leading to their successful application as indicators of human stressors (e.g. Schäfer et al., 2007; Menezes et al., 2010), and giving a high potential for their ecological preferences to be used to predict distribution patterns.

2. Material and methods

2.1. Modelling concept

The HS-MSDMs predict the probability of occurrence (including the probability of detection) of individual taxa that make up the community at individual sampling dates. We assume that there was sufficient time for colonization so that the community reflects the environmental conditions at the site at a time scale of seasons to years. The model describes the response of the community to changes in environmental conditions over these timescales without resolving short-term dynamics (including consequences of individual flood or pollution events). This is compatible with the typical formulation of ecological preferences in databases (representing time averaged responses) and with common monitoring designs that include annual or seasonal sampling and avoid sampling shortly after disturbance events such as floods. Nonetheless, we can include environmental influence factors that change between sampling seasons or years. Inputs to the HS-MSDMs are the regional pool of taxa within the modelled area (derived from biomonitoring data), the environmental influence factors at sampling sites and dates (either measured directly in the field, or modelled from other influence factors), and the ecological preference information (extracted from the ecological databases) of the taxa.

The concept underlying our modelling approach is to use ecological preferences from databases to formulate habitat suitability functions. For each environmental influence factor, the habitat suitability function quantifies on a scale from zero to one how the habitat suitability depends on the taxon and the environmental influence factor. The habitat suitability then serves as direct inputs to the occurrence models instead of using the environmental influence factors directly. As the habitat suitability functions are functions of the ecological preferences and the environmental influence factors, the full model still describes occurrence probabilities of taxa in the community depending on the environmental influence factors. However, the intermediate step of using habitat suitability allows us to use prior information about ecological preferences, which should be beneficial in particular for rarely occurring taxa that are difficult to model without using prior information. In addition, we can check the consistency of the ecological preference information in the database with the observed data and gain hints for improving this information.

In the following subsections, we first describe the construction of the habitat suitability functions (Section 2.2), then the sequence of models analysed in this study (Sections 2.3–2.7), and finally the procedures and measures used for parameter estimation and model evaluation (Section 2.8). Throughout this paper, we use the following indices:

- Sites: \( i \in \{1, \ldots, I\} \)
- Sampling dates at site \( i \): \( t_i \in \{1, \ldots, T_i\} \)
- Taxa: \( j \in \{1, \ldots, J\} \)
- Ecological preferences and corresponding environmental influence factors: \( r \in \{1, \ldots, R\} \)

2.2. Habitat suitability functions

Information regarding ecological preferences extracted from databases (further referred to as ecological preference scores) can be used to parameterize habitat suitability functions which output a taxon-specific habitat suitability (= a normalized and interpolated preference score) as a function of an environmental influence factor (Fig. 1). We can either use the habitat suitability functions parameterized with the fixed ecological preference scores to calculate habitat suitabilities regarding each influence factor for each taxon at each site and date, or we can consider the ecological preferences as uncertain parameters of the habitat suitability function to be updated with independent monitoring data jointly with the other model parameters. In the latter case, we directly enter the habitat suitability function and its ecological preference parameters into the model (rather than calculating the habitat suitabilities as an intermediate step before running the model).

We construct habitat suitability functions, \( h_i(x_{it}, s_j, u_r) \), for each ecological preference, \( r \) (e.g. temperature sensitivity) that takes values normalized between 0 and 1 dependent on the corresponding environmental influence factor at that sampling site and date, \( x_{it} \), (e.g. a temperature of 17 °C), the normalized ecological preference scores of the taxon, \( s_j \), and potentially additional parameters, \( u_r \), that influence the shape of the habitat suitability function for all taxa equally (Fig. 1).

See Appendix A in Supplementary material for further details on the construction of habitat suitability functions, and Appendix B in Supplementary material for further details on habitat suitability normalization. Note that we refer to environment-preference combinations by only referring to the index, \( r \), as each ecological preference is uniquely linked to one corresponding environmental influence factor). The shape of the habitat suitability functions and the number of ecological preference parameters depends on the formulation of the ecological preferences (categorical, continuous, etc.) in the ecological databases.

For habitat suitabilities derived from preference scores that were specified for discrete intervals of a continuous environmental influence factor, we avoided abrupt changes in habitat suitability at class boundaries by approximating a step function by a piecewise linear, continuous function (red dashed lines in Fig. 1A–C).

Evaluating the habitat suitability functions, \( h_i \), at all sites, \( i \), and sampling dates, \( t_i \), for all taxa, \( j \), and for all environmental influence factors, \( r \), leads to a set of habitat suitabilities \( h_{i,j,r} \):

\[
h_{i,j,r} = h_i(x_{it}, s_j, u_r)
\]

(1)

For preference information specified by a classification of taxon into sensitive or insensitive, e.g. to a particular pollutant, we parameterized the habitat suitability for sensitive taxa as a continuous function of the pollutant concentration (the suitability is 1 for insensitive taxa). This leads to the use of additional parameters, \( u_s \), to specify the shape of the function. As an example of such a function, we used an s-shaped function to relate the habitat suitability of sensitive taxa to the insecticide land use index, which serves as a proxy for insecticide pollution:

\[
h_i(x_{it}, s_j, k) = \frac{1}{1 + (x_{it} - s_j \cdot k)^2}
\]

(2)

Here, \( s_j \) is the normalized ecological preference score (0 for insensitive, 1 for sensitive taxa), \( x_{it} \) is the corresponding pollutant proxy, and \( k \) is an additional parameter that determines how fast the habitat suitability decreases with increasing values of \( x_{it} \) (see Fig. 1E and Appendix A1 in Supplementary material). The parameter \( k \) is not taxon-specific; it is part of the parameter vector \( u \) introduced in Eq. (1).

Next, we present our stepwise model development from a knowledge-driven model that integrates ecological preference scores as direct, fixed inputs, to a model that treats ecological preferences as uncertain
parameters to be included in Bayesian inference. We present the different steps to demonstrate different approaches of integrating prior knowledge on ecological preferences into HS-MSDMs with increasing complexity and to assess the tradeoffs between model complexity and gain in ecological understanding and predictive performance. Additionally, for the Bayesian models we evaluate whether prior knowledge is supported by evidence from the data or if a revision of the ecological knowledge would lead to a better agreement of the model output with data.

2.3. Knowledge driven HS-MSDMs (M1 and M2)

These models rely primarily on the prior knowledge about ecological preferences to derive probabilities of occurrence of taxa at given sites and dates. Different model versions use either the habitat suitability to each of the influence factors on their own, or take the minimum or mean across the different habitat suitabilities. Model M1 assumes that the resulting habitat suitability can directly be interpreted as the probability of occurrence of the taxon at the site and date. The M1 version that takes the minimum habitat suitability is described by Eqs. (3)-(5). The random variable \( Y_{i,j} \) describes the occurrence of taxon \( j \) at site \( i \) and date \( t \), \( Y_{i,j} = 1 \) encodes presence of the taxon, \( Y_{i,j} = 0 \) encodes absence; note that we use upper case letters for random variables and the same letter in lower case for its values) given the habitat suitability \( h_{i,j} \) for that taxon at that site and date for all environment-preference combinations \( r \). The probability of occurrence is used to parameterize a Bernoulli distribution (Eq. (5)). Note that to simplify notation, in the following model equations we omit the dependence on environmental conditions and known parameters, but already introduce a vector, \( \theta \), of parameters to be estimated from the observations.

\[
P_{i,j} = \min(h_{i,j} \oplus) \tag{3}
\]

\[
P(Y_{i,j} = 1|\theta) = \begin{cases} 
P_{0} & \text{for } P_{i,j} \leq P_{\min} \\
P_{i,j} & \text{for } P_{\min} < P_{i,j} < 1 - P_{\min} \\
1 - P_{\min} & \text{for } P_{i,j} \geq 1 - P_{\min}
\end{cases} \tag{4}
\]

\[
P(Y_{i,j} = 0|\theta) = \begin{cases} 
1 - P(Y_{i,j} = 1|\theta) & \text{for } Y_{i,j} = 1 \\
1 - P(Y_{i,j} = 1|\theta) & \text{for } Y_{i,j} = 0
\end{cases} \tag{5}
\]

The probabilities were bounded to the interval \([P_{\min}, 1 - P_{\min}]\) by Eq. (4) to avoid data-model conflicts in the cases of observed taxa with \( P_{i,j} = 0 \) or unobserved taxa with \( P_{i,j} = 1 \). For our application, we chose \( P_{\min} = 0.001 \) because all taxa occurred at least once in the data set with 563 samples. Data-model conflicts would lead to a value of zero for the probability of the observed outcome and therefore to a value of zero of the likelihood for the whole community.

Model M1 uses environmental conditions, \( x \), given ecological preference scores, \( s \), and potentially additional parameters, \( u \), as input to the habitat suitability function to calculate all suitabilities, \( h_{i,j} \), according to Eq. (1) and used in Eq. (3), but does not have any parameters to be estimated from the data. This means that the vector of parameters to be estimated for model M1 is empty: \( \theta = \emptyset \). By contrast, in model M2 we introduce taxon-specific factors \( f_{j} \), which make it possible to adjust the probability of occurrence of each taxon individually based on its frequency of occurrence (prevalence) (Eq. (6)). This model requires the calibration of the parameters \( \theta = \{ f_{j} \} \) using independent monitoring data (here and for the estimated parameters of the following models we use curly braces to indicate that this is a set across all values of the indices).
This extension allows us to account implicitly for additional factors that limit the occurrence of taxa (such as dispersal limitations or competitive disadvantages not related to the environmental conditions included in the model). We bounded the values of the factors $f_j$ to the interval $[0,1]$ to avoid probabilities that exceed 1. In both models M1 and M2 we assume the ecological preference score to be 1 if prior information is missing.

2.4. Nonlinear HS-MSDMs with fixed preference scores (M3a and M3b)

Model M3 extends the concept of generalized linear models (GLMs) for the probability of occurrence of taxa by replacing environmental influence factors by habitat suitabilities calculated from environmental influence factors using the habitat suitability functions $h_r$. We use a logistic transformation (Eq. (7), logit link function) of a linear combination of the habitat suitabilities with coefficients $\beta_r$ and $\alpha$ (Eq. (8)):

$$
\text{logistic} = \frac{1}{1 + \exp(-z_{ij})} 
$$

$$
z_{ij} = \alpha + \sum_r \beta_r h_r(x_{ijr}, s_{jr}, u_r) 
$$

The conceptual diagram of this model is presented in Fig. 2. In this model, the additional parameters, $u_r$, in the habitat suitability functions, $h_r$, consists of the parameter $k$ that determines how fast the habitat suitability decreases with increasing pesticide land use index $x$ according to Eq. (2) (see figure in Appendix A1 in Supplementary material) and a weighting factor for the contribution of urban areas use to insecticides pollution $w_{UA}$ (see Appendix C, Table C1 in Supplementary material). They are now also estimated rather than taken as a fixed value. The set of estimated parameters of model M3a is then

$$
\theta = (\alpha, \beta_r, u_r). 
$$

The structure of HS-MSDM M3a is non-linear in its parameters as the suitabilities are not fixed but depend on parameters $u_r$ that are estimated jointly with the other parameters $(\alpha, \beta_r)$ of the model. The model is a multi-species model as the coefficients $\beta_r$ are universal for all taxa and describe how strongly the community as a whole responds to the habitat suitability of each environmental influence factor, $r$, i.e. the parameters $\beta_r$ are inferred from a joint inference to the monitoring data of all taxa. The model is not “joint” in the sense of parameterizing the residual correlation among taxa explicitly (e.g. Ovaskainen and Soininen, 2011; Pollock et al., 2012; Warton et al., 2015). The parameters $\alpha$, which are inferred to adjust the probability of occurrence at all sites, are taxon-specific. Note that the habitat suitabilities are also taxon specific, and thus allow for taxon specific responses to the environmental conditions.

Model M3b further extends Eq. (8) by including a site effect $\gamma$ as an additive term in $z_{ij}$ (Eq. (9)). This term modifies the probability of occurrence of all taxa at the site, and is also called a 'random effect' in generalized linear mixed models (e.g. Bolker et al., 2009, but see Gelman, 2005):

$$
z_{ij} = \alpha + \sum_r \beta_r h_r(x_{ijr}, s_{jr}, u_r) + \gamma_i
$$

$$
\gamma_i \sim N(0, \sigma_\gamma) 
$$

The site effects $\gamma_i$ are normally distributed with a mean of 0 and a standard deviation $\sigma_\gamma$ to be estimated during inference (Eq. (10)). Repeated samples within the same site have the same site effect. The site effect can therefore capture taxon-independent spatial patterns not yet explained by the environmental influence factors in the model that are not dependent on the sampling time. (Instead of a site effect that is constant over time one could also consider a sampling effect $\gamma_{it}$ that varies over sites and sampling times.) The analysis of site effects can
help derive hypotheses on ecological causes for the observed patterns not yet considered in the model. The set of estimated parameters of this model M3b is then \( \theta = [\alpha, \beta, u, \gamma] \).

2.5. Non-linear HS-MSDMs with uncertain ecological preference parameters (M4a and M4b)

The models M4a and M4b are based on the models M3a and M3b, respectively (Eqs. (7)–(9)). However, in the models M4a and M4b all preference scores (now called preference parameters), \( s_j \), are treated as uncertain parameters that are included in the inference. The values from the ecological databases are used as the modes of uncertain prior distributions, whose marginal posterior distributions are then derived by Bayesian inference combining this prior information with evidence from the independent monitoring data (e.g. national invertebrate biomonitoring data in our case study). A subsequent comparison of prior and posterior marginal distributions then quantitatively demonstrates for which parameters we can learn from the monitoring data. As for model M3a and M3b, model M4a does not include a site effect while M4b does. The sets of estimated parameters of these models M4a and M4b are \( \theta = [\alpha, \beta, s_j, u_j] \) and \( \theta = [\alpha, \beta, s_j, u_j, \gamma] \), respectively.

Due to the inclusion of the preference parameters (often multiple parameters per environmental influence factor and per taxon; e.g. Table 1), the number of parameters of this HS-MSDM becomes very large. This does not necessarily lead to identifiability problems as we use prior knowledge about these parameters from the databases and do not leave the model the freedom to deviate strongly from them (see Section 3.3 below for details on priors).

2.6. Hierarchical non-linear HS-MSDMs (M5a and M5b)

M5 models extend M4 models (Eqs. (7)–(9)) with a hierarchical structure in which the taxon-specific parameters \( \alpha_j \) are constrained by an overarching community distribution, which is a normal distribution with mean \( \mu_s \) and standard deviation \( \sigma_s \) that are themselves estimated from the community data jointly with the other model parameters:

\[
\alpha_j \sim N(\mu_s, \sigma_s)
\]

Hierarchical model structures are suggested as an improvement when modelling whole communities because individual taxa (especially rare taxa which often cannot be modelled individually) can borrow information regarding their parameters from the overall community (Warton et al., 2015). Additionally, there is often a lower risk of overfitting individual taxon parameters, which should lead to more reliable predictions (Cardama et al., 2019). Model M5a does not include a site effect (estimated parameters \( \theta = [\mu_s, \alpha_s, \alpha_j, \beta, s_j, u_j] \)) while M5b does (estimated parameters \( \theta = [\mu_s, \alpha_s, \alpha_j, \beta, s_j, u_j, \gamma] \)). M5b is thus the most complex HS-MSDM including the site effect, the uncertain preference parameters, and the hierarchical community distribution of parameters \( \alpha_j \).

2.7. Compilation to community models

We assume that the observations of different taxa at different sites and sampling dates are independent of each other. Therefore, the probability of any outcome, \( y \), (across taxa, sites and sampling dates) is given by the product of the probabilities for individual observations:

\[
P(y|\theta) = \prod_{j=1}^{I} \prod_{i=1}^{L} \prod_{t=1}^{T} P(y_{itj}|\theta)
\]

(12)

In this equation, the probability \( P(y_{itj}|\theta) \) refers to one of the models M1 to M5 introduced in the previous sections. As some of the model equations of the more complex models build on the equations of the simpler models, to clarify the definitions of all models, the complete sets of equations used for each model and the estimated model parameters are listed in Table 1. Note that the Models M3 and M4 use the same equations, but differ in the parameters that are estimated by combining prior information with the observation data.

Bayesian inference will be applied to estimate the parameters of the models M3 – M5. As the priors depend on the case study and, in particular, on the chosen habitat suitability functions, these priors are not part of the model description, but part of the description of the model application (Section 3.3).

2.8. Parameter estimation and model evaluation

When substituting the actual observations into Eq. (12), this expression becomes the likelihood function of the parameters. We estimate the parameters for the model M2 by maximum likelihood parameter estimation. The estimate is then given as those parameter values, \( \hat{\theta} \), for which (the log of) the function \( P(y|\hat{\theta}) \) according to Eq. (12) is maximum for the given observations \( y \). We do not conduct uncertainty estimation of these parameters, as this model only serves as a reference for the more comprehensive models. For the models M3, M4 and M5 we estimate the parameters by combining prior information (described in Section 3.3 below) with evidence from the data and deriving a posterior parameter distribution \( f(\theta|y) \) by Bayesian inference according to:

\[
f(\theta|y) \propto f(\theta) P(y|\theta)
\]

(13)

where \( f(\theta) \) is the prior probability density of the model parameters. We obtain a full posterior distribution including parameter uncertainty. However, some diagnostics are only calculated at the maximum of the posterior.

Numerically, maximum likelihood estimation for the model M2 was done using the function “optim” of R (R Core Team, 2019). Bayesian inference of the parameters of the non-linear models (models M3, M4 and M5) was done numerically using a Hamiltonian Monte Carlo
algorithm implemented in the software Stan interfaced from R through the package rstan (Stan Development Team, 2019, Appendix F in Supplementary material).

All models are evaluated for their model fit during calibration to the whole dataset and for their predictive performance using three-fold cross-validation. For the latter analysis, we randomly split the monitoring data into three parts and tested predictions for each part based on model calibration for the other two parts. We choose three folds to allow for a reasonable representation of rare taxa across the training and testing datasets.

Two performance metrics are calculated for each taxon in each model for both model fitting (training data) and prediction (testing data). The standardized deviance for each taxon, \( d_j \), and for the full, multi-taxa model, \( d \), describes the goodness of the model fit, with smaller values indicating a better fit (corresponding to minus two times the log likelihood divided by the number of available data points):

\[
d_j(\theta) = -\frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \log(P(y_{ij} | \theta)).
\]

\[
d(\theta) = -\frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{t=1}^{T} \log(P(y_{itj} | \theta))
\]

where \( n \) is the number of available data points of taxon \( j \) (either presence or absence) and \( n \) is the total number of available data points across all taxa. The standardized deviance corresponds to the mean square of the residuals of a model with normally distributed errors. The \( D^2 \) statistic

\[
D^2(\theta) = \frac{d(\theta) - d_{null\ model}(\theta)}{d_{null\ model}(\theta)}
\]

quantifies the explanatory power of the influence factors, analogous to the \( R^2 \) statistic in linear regression (Guisan and Zimmermann, 2000). It describes the fraction of the deviance of the null model that is reduced by the model. Its value ranges between minus infinity and 1, with values closer to 1 indicating a higher explanatory power of the influence factors and values below zero representing a poorer fit than the null model. The null model in this context contains only the taxon specific parameter \( \gamma \) and \( \chi \), which are equal to zero and assumes no influence of environmental influence factors. It corresponds to assuming a probability equal to the overall observed frequency of occurrence that is independent of the site. All diagnostics, \( d_j \), \( d \) and \( D^2 \), were evaluated at the maximum posterior parameter estimates for \( \theta \), separately for the calibration and validation data subsets during 3-fold cross-validation and for calibration to the full data set.

3. Model application to macroinvertebrates in Swiss streams

3.1. Monitoring data

We derive presence-absence data for macroinvertebrates from the nationwide Swiss Biodiversity Monitoring program (BDM) overseen by the Swiss Federal Office for the Environment, available from the MIDAT central database (https://midat.cs.cf.ch, last accessed 13 July 2017). The BDM focuses on nationwide biodiversity trends and conducts random sampling along a regular grid across Switzerland. We use data between 2010 and 2015, resulting in 563 sampling dates across 482 unique sampling sites. The dataset contains 245, mostly rare, taxa, with 10 taxa occurring in more than half of the samplings and only 61 taxa in more than 10 % of the samplings (Appendix D in Supplementary material). Taxa have a taxonomic resolution up to species level for Ephemeroptera, Plecoptera and Trichoptera (EPT) and remaining taxa at coarser levels (specifically: two taxa at phylum, one at class, three at order, 67 at family and 25 at genus level). The different taxonomic resolution is defined by the BDM monitoring program and relates among others to the cost and expertise to obtain detailed taxonomic identification. The models treat taxa at different taxonomic levels equally. We do not expect to predict well for all taxa, especially the rare ones, based on environmental influence factors alone. Nonetheless, we include all sampled taxa in order to maintain a full overview of the biodiversity as resolved by the data (Note: some taxa observations are removed due to uncertain species determination within species complexes, see Appendix E in Supplementary material for details on this procedure and on the sample collection within the BDM data).

3.2. Ecological preferences and their link to environmental conditions

Prior information about ecological preferences for macroinvertebrates is extracted from the freshwaterecology.info (Schmidt-Kloiber and Hering, 2015), SPEAR (Liess et al., 2008), and Tachet databases (Tachet et al., 2010), and is independent of the BDM monitoring data used in this study. Ecological preferences are linked to environmental influence factors (Appendix C in Supplementary material) that are expected to have a direct influence on organisms and thus a direct link to ecological preferences. The use of direct influence factors as explanatory variables facilitates biological interpretation and should lead to more universal models than models based on indirect influence factors (Guisan and Zimmermann, 2000). Where possible, we derive direct influence factors from indirect factors (e.g. for temperature, saprobic condition, and flow velocity) because measurements of direct environmental factors are often unavailable for all sites. In cases where the estimation of the direct influence factors was very uncertain, we use indirect factors in the model (e.g. a land-use index as a proxy for insecticide pollution, which is hard to quantify directly due to poorly known inputs, and spatial and temporal heterogeneity). The use of influence factors related to human activities allows us to use SDMs to analyze changes in anthropogenic pressures and to evaluate management alternatives (Guisan et al., 2013). We use five combinations of environmental influence factors with ecological preferences: temperature – temperature preference, flow velocity – current preference, saprobic condition – saprophyte, insecticide land-use index – sensitivity regarding pesticides, and substratum classes – substratum preferences. Saprophy refers to the sensitivity of macroinvertebrates to oxygen limitation resulting from the degradation of organic matter. The insecticide land use index is an estimate of insecticide pollution considering different crops weighted for the average number of insecticide applications and urban areas (see Table C1 in Supplementary material).

Data about ecological preferences for temperature, flow velocity, saprobic condition, and substratum classes is normalized to the range between 0 and 1 (Appendix B in Supplementary material). A piecewise linear, continuous approximation is conducted over multiple discrete classes for temperature, flow velocity and saprobic conditions, and habitat suitability for specific sampling sites and dates are then derived from the habitat suitability function (Fig. 1A–C). The habitat suitability for substratum is determined by a weighted average of the normalized ecological preference scores for the different classes multiplied by the proportion of each class within the site (Fig. 1D). Data for pesticide sensitivity is binary (sensitive – insensitive) and converted into a normalized habitat suitability as a function of the insecticide land use index (Fig. 1E).

3.3. Prior probability distribution

For the \( \alpha \) and \( \beta \) parameters of the models M3-M5 we chose wide priors to primarily learn from the data. Nevertheless, the model formulation given by Eqs. (7) and (8), together with the range of the habitat suitability from 0 to 1 indicate reasonable ranges of these
parameters. We can calculate a range of \( z \)-values \( \Delta z_{90/2} \) for which the probability increases from 5% to 95%. The variation in \( \Delta z_{90/2} \) for each ecological preference \( r \) and across all sites should then not be much greater than that range (the range is sufficiently large to significantly modify the probability for \( z \)-values around 0). To well cover this range, we chose a normal prior with standard deviation equal to \( \Delta z_{90/2}/2 \). As we expect a positive response of the probability of occurrence with increasing habitat suitability, we assume positive prior means of \( \Delta z_{90/2}/2 \).

To avoid “inverse interpretation” of the habitat suitability compensated by negative values of the parameters \( \Delta z_{90/2} \), we truncated the priors of the parameters \( \Delta z_{90/2} \) at zero in those models in which we estimated the ecological preference parameters \( \Delta_n \) (M4a,b, M5a,b). In the model application, we consider five environmental factors and the parameter \( \alpha_j \) must be able to adjust the average probability. As the habitat suitability values are not centered but lead to a positive response in terms of values of \( z \), we chose a normal distribution for the prior of the parameters \( \alpha_j \) with a five times larger, but negative mean, \( -5 \Delta z_{90/2}/2 \), and used half of the absolute mean as the standard deviation.

We include insecticide pollution as an environmental factor which combines the effect of the proportion of urban areas with insecticide application rates (IAR) using an estimated mean weighting factor for urban areas \( w_{\text{UA}} \) of 0.6 (Appendix C in Supplementary material). We choose a lognormal prior with mean and standard deviation equal to 0.6 for this weighting factor. Additionally, the parameter \( K_{\text{min}} \) (see Appendix A in Supplementary material) used for parameterizing the dependence of the habitat suitability on IAR for sensitive species needs to be in the order of 0.1 to have a relevant effect on the outcome (0.1 corresponds approximately to the 50th percentile of the data in the case study). As we are uncertain about its value, we use a lognormal distribution for the inverse of \( K_{\text{min}} K_{\text{max}} \) with mean and standard deviation equal to 10.

As we want to avoid that the site effect \( \gamma \) in the models M3b, M4b, and M5b explains the data rather than the environmental influence factors included in the models, we select an exponential prior for its standard deviation \( \sigma_b \) with a mean equal to \( \Delta z_{90/2}/8 \). The mean of the site effect is fixed at zero (Eq. (10)). For the ecological preference parameters inferred in the models M4 and M5 we choose normal distributions truncated to the interval \([0,1]\) with the mean (of the untruncated distribution) centered at the normalized ecological preference score from the database and a standard deviation of 0.2. This informative prior results in the use of preference parameters similar to those in the database except in cases in which there is strong conflicting evidence in the data. In the model M5 we include a normal prior distribution for the hyper-parameter of the community parameter \( \alpha_i \) with as mean the same prior distribution as for the \( \alpha_i \) in the non-hierarchical model M4. As prior for the standard deviation of the community parameter \( \alpha_i \) we use a lognormal distribution with a mean and standard deviation equal to the standard deviation of the priors of the \( \alpha_i \) parameters in the non-hierarchical model M4. For all models, we assumed the marginal priors of the individual parameters to be independent and used their product as the joint prior of all parameters.

4. Results

4.1. Model fit and predictive performance

The M1 models, which are purely driven by prior knowledge on ecological preferences, fit poorly to the whole dataset as shown by their high standardized deviances (Table 2, Appendix G in Supplementary material). The fit improves when the relative frequencies of occurrence of individual taxa are taken into account in the M2 models. However, considering their \( D^2 \) statistics, they are still slightly worse than the null model, which has a standardized deviance of 0.44. The M1 and M2 models are not considered further in view of their predictive performance. The M3 model, where each habitat suitability factor gets its own taxon-independent weight, and in which two parameters related to the consideration of pesticide sensitivity are estimated, performs better than the null model (Table 2). A large improvement in model fit occurs when treating preference parameters as uncertain inputs and allowing them to shift based on evidence from the data (M4). The extension of adding a site effect always improves model fit (compare a and b versions of the models M3, M4 and M5). Making the model hierarchical by applying overarching community parameters, \( \mu_\gamma \), \( \sigma_\gamma \), which constrain the taxon-specific parameters \( \alpha_i \), does not change the results considerably (compare M4 and M5).

The predictive performance of the models (measured as the mean standardized deviance for the 3-fold cross-validation test datasets) is always slightly poorer than the standardized deviance for model fitting to the whole dataset. This reflects the fact that less data is available for model fitting when conducting 3-fold cross-validation, and that it is harder to predict new data than to fit to calibration data. However, the small difference between fit and validation indicates no serious problem with overfitting. The predictive performance of the models with site effects is slightly better to those without.

4.2. Parameter estimates of habitat suitability-based influence factors

The values of the parameters \( \theta_i \) indicate clearly identifiable (positive) effects of the habitat suitability for the model M3 (Fig. 3, left panel). The parameter values shift when ecological preference parameters are inferred in the M4 models. Specifically, temperature preference gains in importance, followed by saprobity, flow velocity and substratum preference, whereas the sensitivity to insecticide decreases (from a dominant to a typical response similar to other factors except temperature). Parameter estimates of the \( \theta_i \) parameters influence each other. Therefore, small differences between parameters (e.g. between
flow velocity, substratum classes and insecticide land-use index) should not be over-interpreted. The estimates of the $r_i$ parameters remain relatively stable between the non-hierarchical M4 and hierarchical M5 models, as well as between models with and without site effect. Moreover, estimates for all parameters remained stable across the 3-fold cross-validation datasets for all models (see Fig. 3 right panel for model M4; plots of the marginal posterior distribution of all estimated parameters for both training and 3-fold cross-validation in model M4b are presented in Appendix H in Supplementary material).

4.3. Community overview

The M4a model can discriminate presence-absence observations for taxa with a frequency of occurrence down to about 5 % (Fig. 4). The discriminatory power of the M4a (as well as M4b, M5a and M5b) is highest for taxa with an intermediate frequency of occurrence.

Taxa that occur almost everywhere or occur only rarely have modelled probabilities of occurrence that are either always high or always low at all sites (Fig. 4). For these taxa, the M4a model fit is always good, as indicated by the low standardized deviance of individual taxa (Fig. 5). However, for these taxa, the explanatory power of the model is generally poor, as indicated by $D^2$ statistics around 0, because the prediction is not influenced by the explanatory variables.

For example, taxa with a relative frequency of occurrence below 10 % obtained a mean standardized deviance of 0.17 ± 0.14, and mean $D^2$ of 0.12 ± 0.12 for the M4a model fitted to the whole dataset. As taxa become more frequent, the standardized deviance is expected to increase. Yet, some taxa maintain low standardized deviances, indicating a particularly good fit of the model to the distribution of these taxa. An example is *Protonemura lateralis* (frequency of occurrence: 53 %), with a standardized deviance of 0.48, leading to a good representation of its geographic distribution (Fig. 6, Appendix I in Supplementary material provides overview statistics for all taxa). Additionally, for some immediately frequent taxa, the environmental influence factors considered in the model have high explanatory power, as indicated by individual taxa $D^2$ statistics close to 0.5, such as *P. lateralis*. Overall, the model’s main ecological potential lies with taxa with an intermediate frequency of occurrence that have a low standardized deviance and a high $D^2$ statistic.

4.4. Parameter inference for preference parameters

A comparison between prior knowledge about ecological preferences from the databases and posterior distributions from inference of the M4a and M4b models illustrates the potential for improving preference information based on evidence from the data (Fig. 7). An example is *Ecdyonurus helveticus* (relative frequency of occurrence: 31 %), whose distribution is well represented by the M4a model (Fig. 6), including a high explanatory power of the habitat suitability-based influence factors ($D^2 = 0.25$) and a good individual model fit ($d_j = 0.78$). The model results suggest a posterior preference for moderate stream temperatures that is higher than expected based on the prior ecological preference parameter values for this species (Fig. 7). Likewise, the spatial distribution of *Leuctra braueri* (frequency of occurrence: 43 %) is well represented by the model and the influence factors considered, ($D^2 = 0.24$, $d_j = 0.41$, Fig. 6). Model outputs suggest that *L. braueri* is less comfortable in cold and more comfortable in moderate temperatures than described by the prior knowledge (Fig. 7). Vermeiren et al. (submitted) provides further ecological analyses of individual taxa patterns.

4.5. Taxa richness and site effects

The HS-MSDMs including site effects (M3b, M4b and M5b) capture taxa richness better than models without site effects (M3a, M4a and M5a). The model without site effect (M4a) does not resolve the dependence of richness on the influence factors very well (Fig. 8). This leads to an overprediction of richness at sites with low observed richness and a tendency of underprediction at sites with high richness. This is corrected by the site effect (M4b). Nonetheless, this model leads to an overprediction of species level richness.

The site effect is not strongly correlated with any of the
environmental influence factors already included in the model (highest correlation is 0.34 with temperature, Appendix J in Supplementary material), and thus represents other, non-modeled influence factors or ecological processes. The site effect (Fig. 9) leads to a reduced taxa richness in the alpine South to South-East areas and an increased richness in the North and North-West corresponding with the Swiss Plateau.

5. Discussion

SDMs have often been criticized for lacking an ecological basis underlying the modeled patterns (Elith and Leathwick, 2009). By including prior knowledge on ecological preferences linked to environmental variables, the different HS-MSDMs developed in the current study aim to increase the causal and ecological basis of SDMs. The poor results of the M1 HS-MSDM (Table 2) that predicts distribution patterns solely on existing knowledge about habitat suitabilities confirm that models which just consider the fundamental niche are not sufficient to predict species distributions.

Ecological preference scores regarding different environmental factors are often quantified independently of each other and therefore they cannot necessarily be compared in absolute terms. Consequently, simple aggregation rules (such as the minimum or mean in model M1) are likely unsuitable to combine such independently derived habitat suitability factors. Nonetheless, preference scores can provide an elegant way to include niche theory and landscape filter concepts into SDMs. For instance, Bennettsen et al. (2016) combined Habitat Suitability Indices (HSI, trapezoid shaped response curves of individual taxa against environmental variables) as clustered hierarchical sets, representing hierarchical environmental filters where the minimum suitability was taken across clustered sets of HSIs within each hierarchical level. In this case, the HSIs were all derived in a similar fashion and might therefore have been more suitable for combination using aggregation rules than the ecological preferences used in the current study. These results encourage further interaction between (SDM) modelers and developers of ecological databases.

An alternative approach for aggregation is to account for different contributions of habitat suitability based influence factors by including weighting parameters, such as the $\beta$ parameters in the model M3, which describe how strong the community as a whole responds to the habitat suitability of each environmental factor. This improved model fit for the whole community and increased discriminatory ability for taxa with intermediate probabilities of occurrence compared to the models M1 and M2. Similarly, SDMs such as those developed by Pollock et al. (2012) or Ovaskainen et al. (2017) include parameters to relate the influence of traits on the response of taxa to environmental factors in a hierarchical framework. The multivariate nature of these, and our M3 models also, shows the value of accounting for multiple influence factors even if the interest lies in only one specific stressor (Elbrecht et al., 2014).

Existing frameworks that integrate prior ecological knowledge, e.g. regarding ecological preferences or functional traits, to model distributions of taxa, such as RLQ and fourth corner ordinations (Dray et al., 2014), Generalized Joint Attribute Modelling (GJAM, Clark, 2016), Hierarchical Modelling of Species Communities (HMSC, Ovaskainen et al., 2017) or Hierarchical multi-SDMs and mixing models (Pollock et al., 2012, Jamil et al., 2013) have treated prior ecological knowledge as fixed model inputs. In our case, a substantial improvement in model fit occurs with the M4 models, where knowledge on ecological preferences is treated as uncertain priors for model parameters. In fact, model fit during 3-fold cross-validation improves by 17 % and explanatory power increases 3.8 times from model M3 that treats ecological preferences as fixed scores to model M4 that treats them as uncertain parameters. The good convergence of the M4 HS-MSDMs indicates that treating knowledge about ecological preferences as uncertain does not cause identifiability issues for the large number of parameters in our model. Additionally, given the good results, even for taxa with frequencies of occurrence down to about 5 % of samplings with the M4 models (Figs. 4 and 5), HS-MSDMs that include ecological preference as prior parameter values could provide a valuable tool to model less frequently occurring taxa. Note that we cannot expect to
learn much for rare species from the data and that consequently the prior and marginal posterior distribution of the ecological parameters of these taxa are very similar. However, this illustrates the strength of using both prior knowledge and data to improve predictive performance for rare taxa. Existing frameworks could be extended to include uncertain prior ecological knowledge. For instance, GJAM, is a probabilistic framework and would allow for integration of uncertain trait knowledge (Clark, 2016). Likewise, a Bayesian implementation of the HMSC (Ovaskainen et al., 2017) could be extended to also include uncertain prior knowledge regarding the sensitivities of taxa to environmental influence factors. The additional treatment of prior knowledge as uncertain parameters would also benefit decision makers, as it allows us to quantify the uncertainty in available knowledge and model outcomes (Langhans et al., 2018), which is particularly relevant for groups such as macroinvertebrates where knowledge gaps persist and monitoring data can have varying taxonomic resolutions.

**Fig. 5.** Standardized deviance (left panel) and $D^2$ statistic (right panel) versus relative frequency of occurrence for each taxon in the M4a HS-MSDM calibrated to the whole data set. The black line in the left panel indicates the standardized deviance of the null model.

**Fig. 6.** Presence (blue points) and absence (red points) of example taxa (incl. relative frequency of occurrence in percent, $D^2$ and deviance): *Protonemura lateralis* (53%, 0.44, 0.49), *Leuctra braueri* (43%, 0.24, 0.41) and *Ecdyonurus helveticus* (31%, 0.25, 0.78) throughout Switzerland, with the predicted probability of occurrence of the M4a model indicated by the point size. Large blue points and small red points indicate a good model fit.
The integration of prior ecological knowledge into SDMs can be a two-way interaction. HS-MSDMs, such as the M4 models that treat knowledge on ecological preferences as uncertain, have a potential to contribute to complementing and revising ecological preference information. A comparison of the prior (from the database) and posterior (obtained from Bayesian inference with independent monitoring data) preference parameters illustrates the information gained from the data. Particularly, for taxa where the influence factors have a high explanatory power, marginal posterior parameter distributions gained from HS-MSDMs can stimulate the discussion of ecological knowledge with experts to revise preference information in databases and improve models. This discussion should also consider other potential factors that can lead to a disagreement between priors and posteriors, such as the estimation of input variables or model structure deficits. In general, an increase in posterior preference scores compared to the prior provides stronger indication for a need for revision than a decrease, as the latter could be caused by confounding factors not included in the model. An in depth ecological analysis of HS-MSDM outcomes are presented in Vermeiren et al. (submitted).

Especially for the large number of macroinvertebrate taxa with a low frequency of occurrence, other factors than the environmental factors included in the model likely affect their distribution. Such additional processes include dispersal limitation, missing environmental influence factors, neutral processes, density dependent mass effects and biotic interactions (Leibold et al., 2004; Cadotte and Tucker, 2017). For ubiquitous taxa the environmental conditions are obviously not limiting and a null model that predicts a high probability of occurrence everywhere can hardly be improved.

The inclusion of site effects in models M3b, M4b and M5b was aimed at supporting model development and ecological understanding by revealing residual patterns in taxa richness after accounting for the influence factors considered in the model (Fig. 9). In our case study, environmental factors not considered in the model could be increasing the taxa richness in the lowlands. In addition, isolation by distance or harsh environmental conditions due to hydrological disturbance could be hypothesized to explain the lower taxa richness (negative site effect) in the alpine regions. The biogeography of a location as a result of past geological or climatic conditions or the emergence of physical barriers, as well as evolutionary processes affecting genetic differentiation within species influence current day distribution patterns. These factors are not yet included in the model, but could be potential drivers to be considered in further work. In fact, alternative site effects, e.g. representing bioregions, as well as autocorrelation structures could be included in the model to capture spatial processes more explicitly (Domisch et al., 2019, Appendix J in Supplementary material).

In contrast to some mechanistic ecosystem models (e.g. Schuwirth et al., 2016) and SDMs including biotic interactions (for instance as interaction matrices, Kissling et al., 2012), species interactions are not explicitly considered in the current HS-MSDMs although biotic interactions can implicitly affect species responses to environmental factors and thereby affect their realized niches as captured in observational monitoring data. Extending and merging the HS-MSDMs with joint-SDMs that explicitly account for residual correlations among species (e.g. Clark et al., 2014; Pollock et al., 2014; Inoue et al., 2017; Tikhonov et al., 2017; Caradima et al., 2019) would allow exploration of the explicit role of biotic interactions or joint responses to unknown influence factors and may be an interesting line of future research.

The hierarchical structure (M5 model) did not improve model fit or predictive performance (Table 2). This result is likely specific to the well identifiable parameter that was made hierarchical. For parameters that are less well identifiable for each taxon, the gain in information from a hierarchical community structure can be more relevant. For example, including a hierarchical structure to describe spatial or intra-community correlations could be an interesting further step.

Model outputs need to be considered relative to the availability and quality of data on presence-absence, environmental influence factors, and ecological preferences. Even with a huge effort to compile nationwide data on important environmental influence factors, we cannot expect it to be complete. For example, we did not consider seasonal discharge patterns that can affect stream invertebrates and may vary from year to year (Wagner and Schmidt, 2004). Other environmental factors, such as water temperature and insecticide pollution, are modeled using indirect influence factors, because direct measurements were not available across the scale of our study. This introduces additional uncertainty for these influence factors. For example, estimating insecticide pollution from agriculture requires detailed knowledge about
Our results suggest that HS-MSDMs can include uncertain and even uninformative (in case of missing data) prior knowledge on ecological preferences and provide a way to complement and validate current knowledge with independent monitoring data.

Another source of uncertainty comes from the preferences for environmental factors, such as temperature preferences, which are interpolated across different temperature classes, as sharp transitions at the class boundaries are hardly realistic. Additionally, information about ecological preferences for some taxa is incomplete. In fact, information covering a wide range from alpine streams to the lowlands.

Yet, others argue that functional traits and ecological preferences can show a wide diversity even at fine taxonomic levels, thereby limiting the possibility for information from related taxa to fill knowledge gaps (Serra et al., 2016). The HS-MSDM model can assist in this process by interpreted using biological expert knowledge and considering limitations of the model. This fits perfectly in the general process of preference score delineation, in which experts in the field evaluate propositions in an iterative process (Schmidt-Kloiber and Hering, 2015, Serra et al., 2016). The HS-MSDM model can assist in this process by increasing the objectivity of preference score propositions to be evaluated by experts, by generating propositions for taxa about which we have limited prior knowledge, and by evaluating the transferability of current (fixed) preference scores to other regions or future scenarios.

6. Conclusions

We developed a series of habitat-suitability based multi-species distribution models of increasing complexity that aim for strengthening the link between ecological preferences and the occurrence of taxa. Integrating prior knowledge from databases about ecological preferences in SDMs in such a way establishes ecologically interpretable links between taxa and their environment. Moreover, treating prior knowledge as uncertain parameters to be updated with independent monitoring data has the potential of improving prediction in the context of models that explicitly integrate prior knowledge. Indeed, this was the case for our macroinvertebrate case study, allowing predictions even for many rare taxa with frequencies of occurrence down to about 5%. Due to the joint structure of the model and the relatively narrow priors of the preference parameters, the model parameters were well identifiable and the predictive performance during cross-validation was only slightly worse than the fit to the data during calibration.

Integration of prior knowledge into SDMs in a Bayesian framework can lead to a two-way interaction. Prior knowledge from preference databases can improve the transferability of SDMs and a comparison of prior and posterior preference parameters can support the revision of ecological databases. Indeed, model-based testing of ecological preference parameters is a valuable contribution to the continued development of databases on ecological preferences, if results are carefully interpreted using biological expert knowledge and considering limitations of the model. This fits perfectly in the general process of preference score delineation, in which experts in the field evaluate propositions in an iterative process (Schmidt-Kloiber and Hering, 2015, Serra et al., 2016). The HS-MSDM model can assist in this process by increasing the objectivity of preference score propositions to be evaluated by experts, by generating propositions for taxa about which we have limited prior knowledge, and by evaluating the transferability of current (fixed) preference scores to other regions or future scenarios.

Declarations

Peter Vermeiren: Conceptualization; Methodology; Software; Validation; Formal Analysis; Data Curation; Writing – Original Draft; Writing – Review and Editing; Visualization.

Peter Reichert: Conceptualization; Methodology; Software; Validation; Formal Analysis; Data Curation; Writing – Original Draft;
Writing – Review and Editing; Funding Acquisition.
Nele Schwürrich: Conceptualization; Methodology; Validation; Formal Analysis; Data Curation; Writing – Original Draft; Writing – Review and Editing; Visualization; Supervision; Funding Acquisition.

Declaration of Competing Interest
None.

Acknowledgements
This study is part of the AQUACROSS project, funded by European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 642317). We thank the Swiss Federal Office of the Environment (especially Yaël Schindler), the BDM coordination office, Hintermann & Weber, and Info FAUNA, © CSCF Neuenburg (especially Nadine Sarbach-Remund) for access to the data and support. We acknowledge Astrid Schmidt-Kloiber, Patrick Leitner and Wolfram Graf for advice on trait-environment matching, Roï Sierra, Raoul Schaffner, Ruth Scheidegger, Karin Ghilardi, Bogdan Cardamis and Mathias Kueenmer for collaboration and data preparation, Simon Speyer, Christian Stamm and Irene Wittmer for insights in insecticide pollution, and Niklaus Zimmermann, Florian Altermatt, Jakob Brodersen, Chris Robinson and Christian Michél for stimulating discussions.

Supplementary data
Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1106/j.eco.lmodel.2020.108956.

References
Bennetse, 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.
Boiler, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.
Bruggemann, J., 2011. A phylogenetic approach to the estimation of phytoplankton traits. J. Phycol. 47, 52–65.
Cadrin, M., Tacker, G., 2017. Should environmental filtering be abandoned? Trends Ecol. Evol. 32, 429–437.
Cardimba, D., Schwürrich, N., Reichert, P., 2019. From individual to joint species distribution models: a comparison of model complexity and predictive performance. J. Biogeogr. 46, 2260–2274.
Clark, J.S., Siefeld, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecol. Appl. 24, 990–999.
Clark, J.S., 2016. Why species tell more about traits than traits about species: predictive ecological theory for decision support in river management. Ecol. Model. 325, 1–12.
Boiler, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.
Bruggemann, J., 2011. A phylogenetic approach to the estimation of phytoplankton traits. J. Phycol. 47, 52–65.
Cadrin, M., Tacker, G., 2017. Should environmental filtering be abandoned? Trends Ecol. Evol. 32, 429–437.
Cardimba, D., Schwürrich, N., Reichert, P., 2019. From individual to joint species distribution models: a comparison of model complexity and predictive performance. J. Biogeogr. 46, 2260–2274.
Clark, J.S., Siefeld, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecol. Appl. 24, 990–999.
Clark, J.S., 2016. Why species tell more about traits than traits about species: predictive ecological theory for decision support in river management. Ecol. Model. 325, 1–12.
Boiler, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.
Bruggemann, J., 2011. A phylogenetic approach to the estimation of phytoplankton traits. J. Phycol. 47, 52–65.
Cadrin, M., Tacker, G., 2017. Should environmental filtering be abandoned? Trends Ecol. Evol. 32, 429–437.
Cardimba, D., Schwürrich, N., Reichert, P., 2019. From individual to joint species distribution models: a comparison of model complexity and predictive performance. J. Biogeogr. 46, 2260–2274.
Clark, J.S., Siefeld, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecol. Appl. 24, 990–999.
Clark, J.S., 2016. Why species tell more about traits than traits about species: predictive ecological theory for decision support in river management. Ecol. Model. 325, 1–12.
Boiler, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.
Bruggemann, J., 2011. A phylogenetic approach to the estimation of phytoplankton traits. J. Phycol. 47, 52–65.
Cadrin, M., Tacker, G., 2017. Should environmental filtering be abandoned? Trends Ecol. Evol. 32, 429–437.
Cardimba, D., Schwürrich, N., Reichert, P., 2019. From individual to joint species distribution models: a comparison of model complexity and predictive performance. J. Biogeogr. 46, 2260–2274.
Clark, J.S., Siefeld, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. Ecol. Appl. 24, 990–999.
Clark, J.S., 2016. Why species tell more about traits than traits about species: predictive ecological theory for decision support in river management. Ecol. Model. 325, 1–12.
R Core Team, 2019. R: a Language and Environment for Statistical Computing. URL: https://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria.

Schäfer, R.B., Caquet, T., Siimes, K., Mueller, R., Lagadic, L., Liess, M., 2007. Effects of pesticides on community structure and ecosystem functions in agricultural streams of three biogeographical regions in Europe. Sci. Total Environ. 382, 272–285.

Schmidt-Kloiber, A., Hering, D., 2015. www.freshwaterecology.info – an online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. Ecol. Indic. 53, 271–282 Database accessed on 29 March 2016.

Schuwirth, N., Dietzel, A., Reichert, P., 2016. The importance of biotic interactions for the prediction of macroinvertebrate communities under multiple stressors. Funct. Ecol. 30, 974–984.

Stan Development Team, 2019. RStan: The R Interface to Stan. R Package Version 2.19.2. URL: http://mc-stan.org/.

Serra, S.R.Q., Cobo, R., Garça, M.A.S., Dolédec, S., Feio, M.J., 2016. Synthesising the trait information of European Chironomidae (Insecta: diptera): towards a new database. Ecol. Indic. 61, 282–292.

Szöcs, E., Coring, E., Bäthe, J., Schäfer, R.B., 2014. Effects of anthropogenic salinization on biological traits and community composition of stream macroinvertebrates. Sci. Total Environ. 468–469, 943–949.

Tachet, H., Richoux, P., Bournaud, M., Usseglio-Polatera, P., 2010. Invertébrés d’eau douce systématique, biologie, écologie. CNRS éditions, Paris, pp. 607.

Tikhonov, G., Abrego, N., Dunson, D., Ovaskainen, O., 2017. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods Ecol. Evol. 8, 443–452.

Vermeiren, P., Abrantes, K., Sheaves, S., 2015. Generalist and specialist feeding crabs maintain discrete trophic niches within and among estuarine locations. Estuaries Coasts 38, 2079–2082.

Vermeiren, P., Reichert, P., Graf, W., Leitner, P., Schmidt-Kloiber, A., Schuwirth, N., n.d. Confronting existing knowledge on ecological preferences of stream macro-invertebrates with independent monitoring data using a Bayesian multi-species distribution model. Freshwater Science (Submitted for publication).

Wagner, R., 2004. Yearly discharge patterns determine species abundance and community diversity: Analysis of a 25 year record from the Breitenbach. Archiv für Hydrobiologie 4, 511–540. https://doi.org/10.1127/0003-9136/2004/0161-0511.

Warton, D.I., Blanchet, F.G., O’Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K.C., 2015. So many variables: joint modeling in community ecology. Trends Ecol. Evol. 30, 766–779.