Confronting existing knowledge on ecological preferences of stream macroinvertebrates with independent biomonitoring data using a Bayesian multi-species distribution model

Peter Vermeiren[1,3], Peter Reichert[1,4], Wolfram Graf [2,5], Patrick Leitner[2,6], Astrid Schmidt-Kloiber[2,7], Nele Schuwirth[1,8]

[1]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

[2]BOKU: University of Natural Resources and Life Sciences, Institute of Hydrobiology and Aquatic Ecosystem Management, Gregor-Mendel-Straße 33, 1180 Vienna, Austria

E-mail addresses: [3]Present address: Radboud University, Department of Environmental Science, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands, peter.vermeiren@gmail.com; [4]Peter.Reichert@eawag.ch; [5]wolfram.graf@boku.ac.at; [6]patrick.leitner@boku.ac.at; [7]astrid.schmidt-kloiber@boku.ac.at; [8]Nele.Schuwirth@eawag.ch

Received 31 October 2019; Accepted 16 October 2020; Published online XX Month 2021; Associate Editor, Daren Carlisle.

Abstract: A wide knowledge base regarding the ecological preferences of benthic macroinvertebrates is synthesized in public databases. This knowledge can assist in disentangling the influence of multiple environmental factors on the probability of occurrence of macroinvertebrates and in identifying anthropogenic impacts on the macroinvertebrate assemblage. We aimed to examine and extend current knowledge on ecological preferences by confronting it with independent biomonitoring datasets and to assess how the taxonomic resolution of datasets and the prevalence of taxa affects our ability to do so. We used a habitat suitability-based multi-species distribution model (HS-MSDM) and applied Bayesian inference to confront current knowledge (formalized as prior probability distributions) against independent biomonitoring data across rivers in Switzerland. Shifts in the resulting posterior probability distributions relative to the priors indicate a disagreement with the current knowledge of ecological preferences. Ecological preferences for temperature and organic matter had the highest influence on the predicted occurrence of macroinvertebrates in the model, followed by flow velocity, insecticide pollution, and substratum. Three-fold cross-validation tests demonstrated that the HS-MSDM predicted the distribution of taxa with a relative frequency of occurrence between 0.2 and 0.8 considerably better than a model without consideration of environmental factors. However, it was less able to predict the distribution of taxa with a frequency of occurrence <0.1 or >0.9. Nine taxa with a frequency of occurrence between 0.4 and 0.8 were identified as potentially useful bioindicators, given their strong association with the environmental factors in the model. We also identified 29 taxa for which part of the ecological preference data, particularly temperature and flow-velocity preferences, should be re-examined. For river morphology, 18 sensitive and 10 insensitive taxa were identified, although direct and uniquely linked prior knowledge regarding morphology was lacking for all taxa.

Phylogenetically derived information on ecological preferences could be integrated and updated to fill gaps in ecological preference databases. However, the taxonomic resolution of the biomonitoring and ecological preference data plays an important role, as we show by identifying families comprising species that respond differently to environmental factors. These results demonstrate the value of conducting biomonitoring at the most detailed taxonomic level possible.

Key words: ecological niches, habitat suitability, taxonomic resolution, biomonitoring, Bayesian inference

A major challenge in ecology and environmental management lies in disentangling the influence of multiple natural and anthropogenic environmental factors on the composition of communities (Elith and Leathwick 2009, Guisan et al. 2013). The use of existing knowledge on ecological preferences (sometimes referred to as ecological traits and here considered in a broad sense to include ecological preferences or tolerances for natural and anthropogenically influenced environmental factors) can contribute to tackling this challenge. Existing knowledge on ecological preferences describes cause–effect linkages, including non-linear relationships, between the occurrence of taxa and univariate environmental factors (Poff et al. 2006, Menezes et al. 2010). By analyzing spatial patterns in the occurrence of taxa with certain ecological preferences, a general understanding of environmental factors that drive community composition can be achieved (Poff et al. 2006). Such an analysis provides a scientific basis for environmental management across large scales. For example, changes in composition, analyzed in terms of ecological preferences across taxonomically diverse communities, can be used to detect ecological impairments.

Stream macroinvertebrates form a species-rich group that is frequently used as a bioindicator for anthropogenic impacts (Schäfer et al. 2007, Stribling et al. 2008, Menezes et al. 2010, Ruaro et al. 2016). However, macroinvertebrate communities often contain a large number of taxa with a low frequency of occurrence (i.e., spatially rare; Nijboer and Schmidt-Kloiber 2004, Arscott et al. 2006). The spatial rarity of individual species makes it difficult to use biomonitoring data to identify the environmental factors determining their spatial distribution patterns; therefore, our ability to use spatially rare taxa as bioindicators is limited, although spatial rarity may also be an indication of high sensitivity to environmental factors (Cao et al. 2001). A long research history has culminated in a rich knowledge base on ecological

preferences of freshwater macroinvertebrates, available in databases such as Poff et al. (2006), Vieira et al. (2006), Tachet et al. (2010), Schäfer et al. (2011), Schmidt-Kloiber and Hering (2015), and Kefford et al. (2020). Ecological preference data is not consistently available in different geographic locations, but data from existing databases could be combined with independent, local biomonitoring data to increase its information content. For example, when the amount of knowledge that can be gained from biomonitoring data is poor, such as for spatially rare species, combining that data with existing ecological preference data can assist in interpreting occurrence patterns. Such combination can be done by explicitly integrating ecological preference data as a source of prior information into statistical species distribution models (Vermeiren et al. 2020) to strengthen their predictive performance when local biomonitoring data is limited.

Information within ecological databases may be uncertain and, in some cases, incomplete. Knowledge about ecological preferences of taxa in databases is often pooled across a wide range of data sources, including controlled experiments and field observations, through a process of literature synthesis and expert opinion (Schmidt-Kloiber and Hering 2015, Serra et al. 2016). This process often lacks evaluation and validation with independent data not used to construct the databases (but see Kissling et al. 2014 for terrestrial mammals). In this case, using a species distribution model to combine existing knowledge on ecological preferences with independent biomonitoring data in a Bayesian framework provides a systematic methodology to examine existing knowledge on ecological preferences. A comparison of the prior distribution with the resulting posterior parameter distribution indicates if there is disagreement between prior knowledge on ecological preferences and the occurrence of species. Sequential

confrontation of prior knowledge on ecological preferences with new independent biomonitoring data can lead to an iterative learning process (Vermeiren et al. 2020).

Biomonitoring using macroinvertebrates is often conducted at coarse taxonomic levels, such as genus or family level, because of the suggested ecological similarity among taxa at these taxonomic levels and the difficulty in species identification (Dolédec et al. 2000, Poff et al. 2006, Beketov et al. 2009). This mix of taxonomic resolutions leads to some taxa within a single study or database reported at species level and others at genus, family, or even coarser taxonomic levels (Lenat and Resh 2001). Coarse taxonomic resolution limits the information content within biomonitoring datasets, which, in turn, limits the use of such biomonitoring data for environmental management (Schmidt-Kloiber and Nijboer 2004) and its ability to fill gaps in existing knowledge on ecological preferences. Phylogenetic niche conservatism suggests that closely related species are ecologically more similar than expected by simple Brownian evolutionary motion (Losos 2008). Consequently, ecological preferences for taxa missing specific information could potentially be derived from phylogenetically related taxa (Poff et al. 2006, Bruggeman 2011). However, ecological preferences of macroinvertebrates can show a high diversity at fine taxonomic levels (Losos 2008, Graf et al. 2009, Serra et al. 2016). Hence, coarse taxonomic resolution within biomonitoring data likely increases uncertainty when deriving ecological preference information from phylogenetically related species.

In this study, we aimed to examine and extend current knowledge on the ecological preferences of macroinvertebrates by confronting it with independent biomonitoring data. We addressed the following research questions (RQ): RQ1) Which ecological preferences are most important to predict the spatial distribution of taxa within invertebrate assemblages? RQ2) Can we improve existing knowledge on ecological preferences by confronting it with independent

biomonitoring data, and will our ability to improve that knowledge be affected by the prevalence of taxa? RQ3) Can we fill gaps in knowledge on ecological preferences in cases where prior knowledge is not available in databases (a) regarding specific taxa for specific environmental factors within the database and (b) regarding all taxa for an environmental factor that is currently not available within the databases? and RQ4) How does the taxonomic resolution of the biomonitoring data affect inferences about ecological preferences from those data?

METHODS

Invertebrate biomonitoring data

We used data on presence and absence of macroinvertebrate taxa collected between 2010 and 2015 by the Swiss Biodiversity Monitoring (BDM) program overseen by the Federal Office for the Environment and available from the Makroinvertebraten-Datenbank (MIDAT database; http://www.cscf.ch/cscf/Makrozoobenthos/MIDAT, accessed 13 July 2017). The BDM program aims to document nationwide biodiversity trends in Switzerland by using a standardized multi-microhabitat sampling method (IBCH Procedure; Stucki 2010; Appendix S1.1) to conduct sampling at regularly spaced sampling points in rivers close to intersections of a regular grid spread across Switzerland. The dataset contained taxonomic resolution up to species or genus level for Ephemeroptera, Plecoptera, and Trichoptera (EPT), with the exception of the Trichoptera families Hydroptilidae, Goeridae, and Lepidostomatidae (Table 1). Non-EPT taxa were available at family level, except for the Gastropod genus *Ancylus* and 6 taxa with a coarser taxonomic resolution (the class Oligochaeta, orders Hymenoptera, Lepidoptera, and Prostigmata, and the phyla Nematoda and Cnidaria; Table 1). We included all sampled taxa to maintain a full overview of the assemblage.

At some sites, taxa complexes containing multiple species were identified only as a species complex, whereas at other sites the individual species were recorded. We used a 5% rule to resolve such cases of taxonomic mismatches, where if any species within a complex was recorded at the species level within at least 5% of the sites, we kept the species level (otherwise we recorded presence of the associated complex). For sites where the complex was recorded to be absent, we kept the absence records for each of the corresponding species. For sites where the complex was recorded to be present, we made a notation of "not available" for the corresponding species because we could not know if the species were absent or present. For modeling we removed sites for a specific taxon when it was noted as not available, which resulted in a reduced number of presence/absence data points for some taxa.

Data on ecological preferences

Knowledge on ecological preferences for macroinvertebrates used in this study was extracted from the freshwaterecology.info (https://www.freshwaterecology.info, accessed on 29 March 2016; Schmidt-Kloiber and Hering 2015), Spear (http://www.systemecology.eu/indicate/; Liess et al. 2008), and Tachet (available at https://www.freshwaterecology.info; Tachet et al. 2010; Table 2) databases. These databases describe different ecological preferences by assigning affinity scores for individual taxa for different environmental conditions. Hereafter, affinity scores refer to the information as originally presented in the databases, and ecological preference scores refer to the information as we entered and examined it using our model (see next 2 paragraphs). For example, in the Tachet database, affinity scores for flow velocity, ranging between 0 for low and 3 for high affinity, were given within different classes. These classes correspond to specific intervals of standing (<0.01 m/s), low (0.01–0.24 m/s), moderate (0.25–

0.5 m/s), and high (>0.5 m/s) flow velocities and, thus, describe the affinity of a given taxon to the discrete flow-velocity classes (Fig. 1B). Affinity scores in the freshwaterecology.info database contain information primarily at species level. The Tachet and Spear databases contain information at species and coarser taxonomic levels.

We first attempted to exactly match the name of each taxon in the biomonitoring data with the name of a taxon in the databases (which could be at species or coarser taxonomic level, depending on the taxonomic resolution in the biomonitoring dataset). When no exact match was found in the databases, or when affinity scores were lacking for a specific combination of taxon and environmental factor, we derived affinity scores phylogenetically (here approximated as pooling information from taxonomically related taxa). To do so we searched for affinity scores at the genus level and then derived affinity scores for the taxon in question by aggregating information from the other species in that genus. To aggregate information, we took the maximum affinity score for each environmental factor across the species, with the exception of sensitivity to insecticide pollution, where we took the minimum affinity score. This is a conservative approach that overestimates, rather than underestimates, the affinity score for the taxon. When no matches were found at the genus level, we conducted the aggregation at the family level. When no matches were found at the family level, affinity scores were recorded as not available. No affinity scores for any environmental factors were derived for 6 taxa that were identified at taxonomic levels coarser than family level: the class Oligochaeta, orders Prostigmata, Lepidoptera, and Hymenoptera, and the phyla Nematoda and Cnidaria.

Affinity scores derived from the databases can be expressed on different scales depending on the database and environmental factor considered. To standardize the information, we normalized the affinity scores to values between 0 and 1, which we used in the model and

hereafter refer to as ecological preference scores. We normalized values by dividing each affinity score for a given environmental factor by the maximum affinity score for that environmental factor (Fig. 1A–F, Appendix S1.3).

Natural and anthropogenic environmental data

For each of the ecological preferences included in the model, we derived data about the corresponding environmental conditions at each of the sites and sampling dates targeted by the BDM program (Table 2, Appendix S1.2). In addition, we included an integrated assessment of multiple components of river morphology (hereafter morphology) based on the Swiss methods for stream assessment (Liechti 2010) as an environmental factor for which no existing ecological preferences were available in a database. We used this morphology factor as a test case for our procedure to derive ecological preferences in the absence of prior knowledge for all taxa for a given influence factor. Data on substratum and morphology were available for all BDM sites and sampling dates. We calculated 4 other environmental factors, as described in Vermeiren et al. (2020), that were not collected at the specific BDM sites and sampling dates (Appendix S1.2). Specifically we: 1) estimated average flow velocity based on Manning's equation (Cowan 1956), 2) derived temperature with a model independently calibrated with 58 recording stations across Switzerland, 3) estimated saprobic conditions (a factor reflecting water quality related to easily degradable organic substances leading to reduced oxygen conditions for macroinvertebrates) with a model calibrated with water-quality data from 345 stations across Switzerland, and 4) calculated insecticide pollution from agricultural land-use types weighted by the average number of insecticide applications and the fraction of treated wastewater in the river.

The habitat suitability-based multi-species distribution model

Vermeiren et al. (2020) developed a species distribution model, the habitat suitability-based multi-species distribution model (HS-MSDM), that applies Bayesian inference to integrate prior knowledge regarding ecological preferences with independent monitoring data into models. Previously, ecological preference information and its uncertainty had mostly been treated as fixed inputs (Vermeiren et al. 2020). Here, we briefly describe the main characteristics of the HS-MSDM (also see Appendices S1.4–S1.6 and Vermeiren et al. 2020). We use the following indices:

Sites: $i \in \{1, \ldots, I\}$

Sampling dates at site i (subscripted for site i): $t_i \in \{1, \ldots, T_i\}$

Taxa: $j \in \{1, \ldots, J\}$

Ecological preferences: $r \in \{1, \ldots, R\}$

We modeled the probability of occurrence (which includes the probability of detection) for individual taxa making up the assemblage at given sites and sampling dates. We assumed there was sufficient time for the observed assemblages to have responded to changes in environmental conditions at the site at a time scale of seasons to years. Model inputs were: 1) the regional taxa pool based on the list of taxa present within the BDM biomonitoring data, 2) x_{it_ir}, the environmental factors (i.e., the explanatory variables or predictors), at site i and sampling date t_i that are uniquely linked to ecological preference r, and 3) the prior knowledge on the ecological preferences derived from the databases.

The ecological preference scores, $\boldsymbol{s}_{jr}$, combined with the environmental factors, x_{it_ir}, describe a habitat suitability function (Fig. 1A–F):

$$h_{it_ijr} = h_r(x_{it_ir}, \boldsymbol{s}_{jr}, \boldsymbol{u}_r) \quad \text{(Eq. 1)},$$

where $\boldsymbol{u}_r$ represents additional, taxon-independent parameters (in our study we included 2 parameters to describe the response of taxa to insecticide pollution, U_{IAR} and K_{invmax}; Fig. 1E, Table S1.2, Appendix S1.4). Bold lowercase letters refer to vectors that are indexed by subscripts for sites, sampling dates, taxa, and ecological preferences. We can then calculate the habitat suitability score, h_{it_ijr}, for each environmental influence factor *r*, taxon *j*, site *i*, and sampling date t_i, which is a value on a continuous scale between 0 (unsuitable) and 1 (suitable). Such habitat suitability scores can be calculated a priori and then entered in species distribution models (Vermeiren et al. 2020). Alternatively, we included the habitat suitability function, $h_r(x_{it_ir}, \boldsymbol{s}_{jr}, \boldsymbol{u}_r)$, itself into the model. Because the habitat suitability functions can take a non-linear shape, the model can describe a non-linear response to the environmental factors. Furthermore, including the habitat suitability functions into the HS-MSDM leads to a non-linear model regarding the parameters.

To derive predictions regarding the probability of occurrence of each taxon *j*, site *i*, and sampling date t_i from the HS-MSDM model (Fig. 2), we calculated a linear predictor, z_{it_ij}, which is a weighted sum over all habitat suitability functions:

$$z_{it_ij} = \alpha_j + \sum_r \beta_r \, h_r(x_{it_ir}, \boldsymbol{s}_{jr}, \boldsymbol{u}_r) \quad \text{(Eq. 2)},$$

where α_j is a taxon-specific parameter that can increase or decrease the probability of occurrence of a specific taxon at all sites and sampling dates and, thus, relates to its overall prevalence across Switzerland. β_r are the taxon-independent weighting factors (with values between 0 and infinity) that apply to the whole assemblage. The β_r parameters describe how strongly the habitat suitability function regarding each environmental influence factor affects the occurrence of taxa within the community. The final term, $h_r\left(x_{it_ir}, \boldsymbol{s}_{jr}, \boldsymbol{u}_r\right)$, is the habitat suitability function from Eq. 1.

To convert the predictor z_{it_ij} in Eq. 2 to probabilities of occurrence $P\left(Y_{it_ij} = 1|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{u}\right)$ between 0 and 1, we applied a logistic transformation:

$$P\left(Y_{it_ij} = 1|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{u}\right) = \frac{1}{1 + \exp\left(-z_{it_ij}\right)} \quad \text{(Eq. 3)},$$

where Y_{it_ij} takes a value of 1 for occurrence and 0 for absence of taxon *j* at site *i* and sampling time t_i, $\boldsymbol{x}$ are the environmental factors, and $\boldsymbol{s}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{u}$ are the model parameters.

Parameter inference

We can assume that the observations of different taxa at different sites and sampling dates are independent of each other. Consequently, the probability of any outcome is given by the product of the probabilities for individual observations:

$$P(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i=1}^{I} \prod_{t_i=1}^{T_i} P\left(y_{it_ij}|\boldsymbol{\theta}\right) \quad \text{(Eq. 4)},$$

where $\boldsymbol{\theta}$ stands for the model parameters: $\boldsymbol{x}$, $\boldsymbol{s}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{u}$. Hence, the probability of occurrence for a specific taxon at a specific site and sampling date (y_{it_ij} = 1) is given by $P\left(Y_{it_ij} = 1|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{u}\right)$, and the probability of absence (y_{it_ij} = 0) is given by $1–P\left(Y_{it_ij} = 1|\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{u}\right)$. When we insert the actual observations from the BDM biomonitoring data in Eq. 4, it becomes the likelihood function for the model parameters given the data. The likelihood function describes how likely the observed data are if the model were true. By searching for model parameter values that lead to the largest likelihood, one can obtain the best agreement between model output and observations (i.e., maximum-likelihood parameter estimation). Additionally, we can account for prior knowledge about the parameters by applying Bayesian inference.

Bayesian inference is well suited to confront existing knowledge (termed prior knowledge) about model parameters (e.g., parameters that describe the ecological preference

scores) with independent data through the process of model calibration. The prior knowledge about the parameters is formulated as a probability distribution. This distribution is wider with weaker prior knowledge and narrower if the prior uncertainty about the parameters is small. The calibration procedure tries to find the best compromise between the prior knowledge and a good fit to the data, which is formalized in a likelihood function. In combination with the prior probability distributions of the model parameters, $f(\boldsymbol{\theta})$, the joint posterior parameter distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$ can then be obtained according to Bayes' theorem:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta})\, P(\boldsymbol{y}|\boldsymbol{\theta}) \quad \text{(Eq. 5)}.$$

The resulting posterior distribution of a specific parameter can be compared with the corresponding prior distribution. If the prior knowledge is confirmed by the data, then the width of the posterior distribution will be narrower than the prior. If there is contradicting information between the prior knowledge and the data, the mode of the posterior will be shifted compared with the mode of the prior (prior-to-posterior shift; Fig. 3). In the case of low information content in the data about the parameters, the prior and posterior distributions will be similar.

For our model, including the habitat suitability functions allowed us to infer the parameters of these habitat suitability functions (including the parameters for the ecological preference scores, $\boldsymbol{s}_{jr}$) during model calibration to the BDM monitoring data. By including the prior knowledge on the ecological preferences, we can systematically examine differences between the marginal prior and posterior probability distributions (including the prior-to-posterior shift and the uncertainty in the probability distributions; Fig. 3) to assess if there is agreement between the information in the databases and the biomonitoring data.

The posterior probability distribution of all parameters combined is referred to as the joint posterior distribution and includes information about the correlation among parameters,

whereas the distribution of each individual parameter is called a marginal posterior distribution. We formulated marginal prior probability distributions for the ecological preference scores $\boldsymbol{s}_{jr}$ in the form of a normal distribution truncated to the interval [0, 1] with a mean centered at the normalized ecological preference scores and with a standard deviation of 0.2 (Fig. 3; Appendix S1.5). We chose this standard deviation to account for the uncertainty of the information in the databases and to allow the posterior parameter distribution to shift during Bayesian inference in case of strong evidence in the BDM monitoring data. For taxa with missing affinity scores, we chose uniform prior distributions, which give an equal probability for all values between 0 and 1 (Appendix S1.5). We used wide prior distributions for both α_j and β_r parameters to allow the posteriors to be shifted by learning from the BDM biomonitoring data during Bayesian inference (Appendix S1.5).

For all numerical simulations (Appendix S1.6), we used Hamiltonian Monte Carlo simulation implemented in the *rstan* package (version 2.19.2) in R statistical software (version 3.5.2; R Project for Statistical Computing, Vienna, Austria). The Stan code of the model is provided in Appendix S1.8.

Model evaluation

We evaluated the HS-MSDM for its model fit to the whole dataset and for its ability to predict using 3-fold cross validation. We chose 3 folds to allow for a reasonable representation of rare taxa across the training datasets. We calculated 3 evaluation metrics: standardized deviances (d), the D_j^2 statistic, and the area under the receiver operating characteristic curve (AUC; Appendix S1.7). Smaller standardized deviances indicate better fit (when evaluated for the whole dataset) or better predictive performance (when evaluated for the 3-fold cross-validation

datasets), and we calculated standardized deviances for each taxon (d_j) and across all taxa (d). The D_j^2 statistic quantifies the explanatory power of the environmental factors for taxon j, with values close to 1 indicating a high explanatory power and values close to 0 indicating a low explanatory power. The AUC is a popular metric used to assess the performance of species distribution models (Jiménez-Valverde 2012). It is based on a comparison of the true positive vs the false positive rate, with a value near 0.5 indicating no ability to separate presence from absence and a value of 1 indicating perfect separation (however, see limitations of this measure as summarized by Lobo et al. 2007). We assessed AUC for each taxon. Note that each of these 3 evaluation metrics are affected by the relative frequency of occurrence (prevalence) of the individual taxa. We analyzed all evaluation metrics at the maximum posterior parameter estimates.

Model application

To address RQ1, we quantified the relative importance of the different ecological preferences on the probability of occurrence of taxa within macroinvertebrate assemblages across all model applications described below. Specifically, we compared the posterior distributions of the β_r parameters for the different ecological preferences. The β_r parameters reflect the influence of each environmental factor as well as the distribution of ecological preferences among taxa (i.e., if all taxa have a similar preference for a given environmental factor, its importance in governing the composition of the assemblage will be less than if taxa had marked differences in their ecological preferences).

We applied the HS-MSDM to the BDM dataset at its finest-available taxonomic resolution (Table 1, dataset S) with stream temperature, flow velocity, saprobic conditions,

insecticide pollution, and substratum classes as environmental factors (MS1; Table 3). Additional hydro-morphological influence factors were considered but not included in the final model because of data scarcity or low sensitivity of the model (see Appendix S1.2). We then identified taxa for which specific preference information should be revised by experts (RQ2) by identifying shifts of >0.2 from the maximum prior to the maximum posterior probability distributions of ecological preference scores (prior-to-posterior shift; Fig. 3). We chose a threshold of 0.2 to focus on taxa with a considerable change for which databases might most benefit from updating in consultation with ecological experts.

We also used model MS1 to examine the ability of the HS-MSDM to derive ecological preference scores for taxa (RQ3a) and for an environmental factor (RQ3b) with missing prior information. To answer RQ3a, we pooled information from the biomonitoring data for phylogenetically related taxa as described previously. We identified taxa with pooled ecological preference scores for which we obtained better model performance (i.e., d_j and D_j^2 statistics) than expected based on their frequency of occurrence (Fig. 4A–D). To answer RQ3b, we attempted to infer ecological preference scores for the environmental factor of morphology from the monitoring data without using prior knowledge. Although the databases contain some prior knowledge for individual aspects of morphology, there is no direct and uniquely linked ecological preference that corresponds to the morphological assessment used in Switzerland. We applied the HS-MSDM with morphology instead of substratum (MS2; Table 3) and with both morphology and substratum (MS3; Table 3) as environmental factors. We classified taxa with a relative frequency of occurrence >0.1 (because the model performs more poorly for rare taxa) and a D_j^2 >0.2 (indicating a reasonable explanatory power) as sensitive or insensitive regarding

morphology when their maximum posterior preference score was >0.55 and <0.45, respectively (indicating a considerable difference from 0.5).

To evaluate the effect of taxonomic resolution (RQ4), we derived a 2nd dataset (Table 1, dataset F) by pooling the EPT species and genera of dataset S at family level and keeping the other taxa at family or coarser levels as in dataset S. We applied the HS-MSDM model to dataset F with stream temperature, flow velocity, saprobic conditions, insecticide pollution, and substratum classes as environmental factors (MF1; Table 3). We then compared the explanatory power (D^2 statistic) of the HS-MSDM when applied to the MS1 and MF1 datasets and qualitatively compared ecological preference scores obtained for pooled families in dataset MF1 compared with those of the individual taxa in dataset MS1.

RESULTS

RQ1—Relative influence of ecological preferences

The different HS-MSDM models had only a slightly higher deviance for cross validation than for calibration (Table 3), indicating a reasonable predictive performance and no issue of overfitting. Ecological preferences related to temperature, followed by saprobic condition, had the highest influence on the probability of occurrence of macroinvertebrates across Switzerland in all models, with the other environmental factors following but varying in their order of influence among models, as indicated by the posterior distributions of the β_r parameters (Fig. 5). All ecological preferences contributed to explaining the observed distribution patterns, as indicted by the positive posterior distributions of their β_r parameters that did not overlap with 0 (Fig. 5). Exceptions were substratum and morphology in model MS3, which had lower β_r parameter values that overlapped with 0, suggesting some redundancy in the information content

of substratum and morphology. However, the Pearson's correlation coefficients between morphology and each of the 7 substratum classes were low (mean: –0.07 ± 0.17 SD, range: –0.23–0.28).

RQ2—Confronting existing knowledge of ecological preferences with data

By comparing prior and posterior distributions of the preference parameters, we were able to identify taxa and preferences for which inferred information from the monitoring data contradicts existing knowledge. However, for this comparison it is important to consider the goodness of fit (deviance, d_j) and explanatory power (D_j^2) of the model, which are affected by the frequency of occurrence of the taxa (Fig. 4A, B). For example, taxa with a very high (>0.9) or low (<0.1) frequency of occurrence often obtained a deviance below 0.2 (Fig. 4A, Appendix S2), but the model's explanatory power was often limited in these cases (Fig. 4B) because these taxa were predicted to have a high or low probability of occurrence everywhere. By contrast, the ability of the model to distinguish presence from absence at sites was highest for taxa with an intermediate relative frequency of occurrence of roughly between 0.2 and 0.8 and even including some taxa with a relative frequency of occurrence down to 0.1. Notable examples for taxa with good predictive performance, as evidenced by low deviances (Fig. 4A), in the MS1 model and a relative frequency of occurrence between 0.4 and 0.8 included 2 families, Gammaridae and Elmidae, and 7 species, *Baetis rhodani*, *Baetis alpinus*, *Rhyacophila tristis*, *Protonemura lateralis*, *Drusus discolor*, *Leuctra braueri*, and *Nemoura mortoni*. These taxa also had relatively high D_j^2 values in the MS1 model (Fig. 4B), indicating that the included environmental factors were strongly associated with occurrence of these individual taxa. The MS1 model performance

was comparable across Ephemeroptera, Plecoptera, and Trichoptera, especially when considering species with a relative frequency of occurrence between 0.2 and 0.8 (Fig. 4C, D).

Some ecological preferences shifted in prior-to-posterior distributions when confronted with biomonitoring data. Temperature preferences of individual taxa displayed the largest prior-to-posterior shifts (especially the cold and moderate classes, which are well represented in the monitoring data; Fig. 6A), followed by flow-velocity preferences (particularly low-, moderate-, and high-velocity classes; Fig. 6B) and the beta-saprobic condition class (Fig. 6C). By contrast, there were few prior-to-posterior shifts for insecticide pollution (Fig. 6D) and substratum classes, except for the class of pebbles (Fig. 6E). Ecological preferences did not substantially shift for environmental factors with poor coverage in the monitoring data, such as classes of standing water and xeno- and poly-saprobic conditions.

We generally confirmed the Spear classification of taxa's sensitivity or insensitivity to insecticide pollution (Liess et al. 2008), with only a few observed prior-to-posterior shifts. For example, with model MS1, the genus *Amphinemura*, family Athericidae, and species *Baetis alpinus* obtained narrower marginal posterior probability distributions for sensitivity to insecticide pollution compared with the priors, confirming their high sensitivity. Likewise, the insensitivity of 13 families: Asellidae, Cordulegastridae, Dugesiidae, Elmidae, Erpobdellidae, Gammaridae, Glossiphoniidae, Hydrobiidae, Lymnaeidae, Physidae, Sphaeriidae, Stratiomyidae, and Tabanidae were confirmed by the model. For the suborders Oligochaeta and Prostigmata and the phylum Nematoda, for which no prior knowledge was available, model calibration suggested low sensitivity to insecticide pollution. The largest prior-to-posterior shift occurred for the caddisfly genus *Tinodes* (relative frequency of occurrence: 0.12; insecticide pollution shift: –0.29).

Prior-to-posterior shifts differed in magnitude among taxa depending on their frequency of occurrence. Infrequently occurring taxa often had marginal posterior distributions for preference parameters that were the same as the priors, which was likely a result of the limited information content in the biomonitoring data for rare taxa. For many of the more frequently occurring taxa, the standard deviation of the marginal posterior probability distributions decreased, suggesting that the current available knowledge on ecological preferences was confirmed by independent biomonitoring data (Appendix S3.1). For taxa with a relative frequency of occurrence between 0.2 and 0.8, prior-to-posterior shifts for most ecological preferences were larger compared with all other taxa (Fig. 6A–E). Two examples of taxa with large prior-to-posterior shifts were *Nemoura minima* and *N. mortoni*, species with a relative frequency of occurrence of 0.31 and 0.50, respectively. Their general distributions across Switzerland were represented well by model MS1 (Fig. 7A, B), as also evidenced by the reasonable explanatory power of the environmental factors (D_j^2 = 0.37 and 0.33, respectively) and model fit (d_j = 0.50 and 0.59, respectively). Species-specific temperature preferences were available in the databases and, therefore, included in the model. The slightly narrower marginal posterior distribution compared with the prior confirmed the low suitability of very cold temperatures for *N. minima* (Fig. 7A). By contrast, prior-to-posterior shifts suggested a higher suitability of moderate temperatures for *N. minima* (Fig. 7A) and of very cold conditions for *N. mortoni* (Fig. 7B). Likewise, a narrower posterior distribution confirmed the high suitability of moderate flow velocities for *N. minima*, despite this preference being phylogenetically derived from other *Nemoura* species, but a prior-to-posterior shift suggested a lower suitability of high flow velocities.

Based on our results, we compiled a list of 29 taxa with a frequency of occurrence ≥0.15 for which model MS1 had reasonable explanatory power (i.e., $D_j^2 \geq 0.2$) and that displayed prior-to-posterior shifts exceeding 0.2 for a specific ecological preference, mainly temperature and flow-velocity preferences but also saprobity, substratum, and 1 case of insecticide preferences (Table S3.2).

RQ3—Filling knowledge gaps

We were able to use the model to infer preference information for taxa with missing prior information (RQ3a). In addition to the 6 taxa at taxonomic levels coarser than family level, only few, primarily rare, taxa that were missing prior knowledge about ecological preferences were included in the model, assuming a uniform prior distribution (Table 4). However, many taxa with ecological preference scores that were derived from related taxa at genus or family level were included in the model (Table 4). The average relative frequency of occurrence of taxa with derived ecological preference scores in MS1 was below 0.2 for all ecological preferences, although some taxa had high relative frequency of occurrence (>0.5). One such example is *R. tristis*, which obtained reasonable d_j (0.68) and D_j^2 (0.25) statistics indicating good model fit and explanatory power given its relative frequency of occurrence of 0.55 in model MS1, but which derived all of its ecological preference scores from aggregated information at the genus level. However, prior-to-posterior shifts suggested lower values for cold and moderate temperatures and higher values for warm temperatures (Fig. 7C), indicating that this species differs in temperature preferences from other *Rhyacophila* species. Other non-rare taxa with reasonable model fit and explanatory power also derived most of their ecological preference scores (except temperature preferences) from related taxa at the genus level. These taxa include: *B. rhodani*

(however, we did not derive ecological preferences on insecticide pollution), *P. lateralis*, *N. mortoni*, and *L. braueri* (relative frequency of occurrence: 0.67, 0.53, 0.50, and 0.43, respectively; Appendix S2.1).

We were able to use models M2 and M3 to successfully infer preferences for the environmental factor morphology without using prior information (RQ3b). We classified 18 taxa as sensitive to morphological conditions and 10 taxa as insensitive (Table 5) based on the criterion that their mean posterior ecological preference score for morphology was >0.55 or <0.45, respectively. For the other taxa, the marginal posterior probability distribution was very similar to the prior or the model performance was considered inadequate to make a strong statement on morphological preferences (i.e., for rare taxa with a relative frequency of occurrence <0.1 or for which the explanatory power, D_j^2, of the model was <0.2).

RQ4—Effects of taxonomic resolution

The marginal posterior probability distributions of ecological preference scores can differ between families and their corresponding genera and species, as shown by comparing models MF1 and MS1. For example, the marginal posterior probability distributions inferred from model MF1 for the family Baetidae suggested a high preference for many of the environmental conditions (exceptions were the very cold temperature class, the xeno- and poly-saprobic classes, and the substratum classes mud, roots-litter, and microphytes; Fig. 8). By contrast, some of the individual species or genera obtained markedly different ecological preference scores, as indicated by their marginal posterior probability distributions inferred with model MS1 (Fig. 8). For example, the results suggested low preferences of *Baetis lutheri* for very cold and cold water temperature classes and of *B. alpinus* for warm water temperatures in contrast with the high

preference across temperature classes at the family level. The information content lost from pooling individual species to family level was also reflected in the lower explanatory power of model MF1 as a whole compared with model MS1 (D_j^2 = 0.148 and 0.183, respectively; Table 3). Moreover, model MF1 had lower β parameter values for temperature and flow velocity and higher values for substratum than model MS1, indicating that species (EPT in the BDM dataset) differ in their sensitivity to these environmental factors compared with their corresponding families.

DISCUSSION

This study demonstrated that the integration of prior knowledge and independent biomonitoring data in an HS-MSDM can lead to realistic predictions of macroinvertebrate occurrences for many taxa with frequencies of occurrence between 0.2 and 0.8. We took advantage of the information content within both data sources to examine and extend current knowledge on ecological preferences. Both the frequency of occurrence of taxa and taxonomic resolution of the data played important roles in the ability to update prior knowledge on ecological preferences through confrontation with independent biomonitoring data. For rare taxa, there was limited information (i.e., few presence data points) in the biomonitoring data. Hence, the ability to improve predictions on the occurrence of rare species by integrating prior knowledge is particularly valuable because these species could be useful bioindicators when their spatial rarity is linked to preferences for specific environmental factors. Likewise, the unique ecological preferences of individual species deserve close attention in biomonitoring programs, which our study highlights by showing improved predictions with increased taxonomic resolution.

RQ1—Relative influence of ecological preferences

Species distribution models often use a set of variables that simplify reality to describe and predict spatial patterns (Elith and Leathwick 2009, Guisan et al. 2013). Previous studies have aimed to increase the realism and reliability of such models by including processes such as biotic interactions, dispersal limitations, or temporal dynamics (Guisan and Zimmermann 2000). In the current study, we aimed to increase the mechanistic foundation of species distribution models by explicitly including relationships between species occurrences and ecological preferences for several environmental factors. This was done by including habitat suitability functions into the structure of the HS-MSDM and integrating existing knowledge on ecological preferences of individual taxa as prior parameter values for these habitat suitability functions.

The importance of each environmental factor included in the HS-MSDM model is affected by the range of the environmental conditions covered by the monitoring data as well as the strength of the taxa's ecological preferences for each environmental factor. For instance, rivers sampled in the BDM program are spread across a broad elevation gradient (minimum = 200 m a.s.l., maximum = 2630 m a.s.l.), corresponding to a large temperature range. This temperature gradient offers a wide niche axis along which taxa could diverge, and, indeed, ecological preferences for temperature had the highest influence in explaining the observed occurrence patterns. Associations between temperature and spatial patterns in macroinvertebrate taxa relative abundance and density have also been observed at a smaller scale within individual river catchments in the Swiss Plateau (Robinson et al. 2014), further suggesting an influence of temperature on assemblage structure of lotic macroinvertebrates. Ecological preferences for saprobic conditions, the 2nd-most influential environmental factor in the model, confirm that

many taxa display strong ecological preferences for specific saprobic conditions (Schmidt-Kloiber and Hering 2015) and that saprobic conditions play an important role in determining the composition of macroinvertebrate assemblages. This ecological preference is influential despite improvements in reducing organic matter pollution in rivers in Switzerland (Hering et al. 2012), and few of the BDM sampling sites fell in the extreme xeno- and poly-saprobic classes.

RQ2—Confronting existing knowledge of ecological preferences with data

The HS-MSDM provides a systematic framework that can be used to confront available information on ecological preferences of individual taxa with independent data. A comparison of the marginal prior and posterior parameter distributions for taxa with sufficient model performance can help identify taxa for which experts should consider revising preference information in databases. We propose a 2-step process: 1) model cross validation, which tests the predictive performance and explanatory power of the model, and 2) ecological expert review of the proposed changes, considering knowledge about the ecology of the taxa and the inherent uncertainty within the modeling process (Vermeiren et al. 2020).

Including prior knowledge on ecological preferences into species distribution models can lead to good predictive performance, even for relatively rare taxa (Vermeiren et al. 2020). Results from our model cross validation confirm a good predictive performance for many taxa with a relative frequency of occurrence between 0.2 and 0.8 and even for some taxa with a relative frequency of occurrence down to 0.1 (Fig. 4A–D). The biomonitoring datasets we used contained a large number of spatially rare taxa (65% of taxa in dataset S were present in <5% of the samples), as is often the case for macroinvertebrate assemblages (Nijboer and Schmidt-Kloiber 2004, Arscott et al. 2006). Because only few presence data points are normally available

for rare taxa, biomonitoring data do not contain enough information to infer a narrow marginal posterior probability distribution, especially when prior knowledge is also lacking. The moderate explanatory power of the model across the whole assemblage reflects, in part, the difficulty of representing the distribution of rare taxa.

A prior-to-posterior shift towards increased ecological preference scores (or lowered sensitivity for insecticide pollution) suggests that a taxon can occur under environmental conditions that were thought to be unsuitable and provides an indication to reconsider current knowledge. By contrast, a shift to lowered ecological preference scores (or higher sensitivity for insecticide pollution) may also be the result of confounding factors, which could restrict a taxon from occurring under certain combinations of environmental conditions despite its tolerance to a specific condition. Overall, in our study, we learned most about preferences for temperature and flow velocity, as indicated by the higher average prior-to-posterior shifts of these ecological preferences compared with others (Fig. 6A–E).

Our results indicate that the Spear classification of species being sensitive or insensitive to insecticide pollution, used for bioindication with the Spear index (Liess et al. 2008), is a reliable source of prior information to predict the distribution of taxa. We observed only few prior-to-posterior shifts in ecological preference for insecticide pollution and instead often observed a reduced uncertainty of the ecological preference knowledge (i.e., a narrower posterior probability distribution compared with the prior). The few large prior-to-posterior shifts we observed were for taxa that lacked species-specific knowledge. In such cases our model might offer an opportunity to learn about sensitivity to insecticides. For example, the large prior-to-posterior shift observed for the caddisfly genus *Tinodes*, classified as sensitive in the Spear database (Liess et al. 2008), may imply either that some species of this genus are less sensitive

than others or that this genus is generally less sensitive to insecticides than previously thought. Similarly, for the caddisfly species *Ecclisopteryx madida* (relative frequency of occurrence: 0.11; insecticide pollution shift: –0.10), the Spear classification was derived from other species of the same family because of missing species-specific information. The lowered sensitivity suggested by the HS-MSDM indicates a difference in the sensitivity of this species compared with other taxa of the same family, which consists of a large number of different genera and species. More accurate data regarding insecticide pollution, instead of the currently available rough estimate based on wastewater and agricultural land-use inputs, would further increase our ability to learn from the inference.

Based on our results, we compiled a list of 29 taxa (Appendices S3.2, S3.3) for which we suggest that ecological preference information in the databases should be reexamined by experts because these taxa displayed large prior-to-posterior shifts. Such shifts may illustrate imperfect information in the ecological preference databases or show the effects of deriving information from phylogenetically related taxa that do not have exactly the same preferences. In both cases, we recommend a revision or extension of ecological preference information.

RQ3—Filling knowledge gaps

Despite the high cost and required level of expertise, conducting biomonitoring and synthesizing ecological knowledge in databases at the most detailed taxonomic level possible is highly valuable for improving our ecological understanding and predictive capacities. For example, for *R. tristis* and *Rhyacophila hirticornis* (relative frequency of occurrence of 0.55 and 0.22, d_j of 0.25 and 0.40, D_j^2 of 0.68 and 0.27, respectively), we derived prior knowledge on their ecological preferences by pooling scores from taxonomically related species within the genus

Rhyacophila. Nonetheless, both species showed large prior-to-posterior shifts for some ecological preference scores, which differed between the 2 species (Fig. 7C, D). For example, the prior-to-posterior shifts suggested higher ecological preference scores for warm (*R. tristis*) and very cold (*R. hirticornis*) temperatures as compared with the prior distributions that were phylogenetically derived.

The HS-MSDM can be applied to derive ecological preference scores when prior information is lacking for all taxa in the model. Model MS2, for example, obtained good overall performance and included morphology, an environmental factor for which no single, uniquely linked ecological preference was available (Table 5). Moreover, results of models MS2 and MS3 indicated that 18 sensitive taxa, especially those with relatively high relative frequency of occurrence, such as *B. alpinus*, *Baetis muticus*, *D. discolor*, *L. braueri*, *N. mortoni*, and *P. lateralis*, can be expected to respond to morphological alterations. This information could be useful for monitoring the success of restoration measures, although the multiple criteria integrated into the Swiss morphological assessment might obscure a direct causal identification of why a species might be more or less sensitive.

RQ4—Effects of taxonomic resolution

Because of resource constraints, a lack of taxonomic expertise at species level for some taxa (including missing determination keys for species-level identification), or the lower error rates that can be achieved for coarser-level identification, monitoring is sometimes conducted at a coarse taxonomic resolution (Stucki 2010). Nonetheless, our results for EPT taxa at family-lever or coarser resolution gave poorer model fit and predication than results at species level (Table 3; MF1 vs MS1). These results confirm that ecological preferences can be variable among

species within the same genus or family and that the information content of biomonitoring data is highest at the finest taxonomic resolution (Schmidt-Kloiber and Nijboer 2004, Serra et al. 2016).

For taxa where species-level data on ecological preferences is already available, such as the EPT taxa in this study, our modeling approach can help identify specific bioindicator species within a family. This was the case for common families that consist of species that differ in their ecological preferences including Baetidae, Heptageniidae, Limnephilidae, Nemouridae, Perlodidae, Rhyacophilidae, and Taeniopterygidae. For common taxa that are currently resolved to a coarser level than family in the BDM dataset, such as nematodes and oligochaetes, the model has a low explanatory power, even though it is known that these taxa can be useful bioindicators at a finer taxonomic resolution (e.g., Vivien et al. 2016, Höss et al. 2017). Also, for common non-EPT families, such as the dipteran families Chironomidae, Limoniidae, Simuliidae, and Empididae, a finer taxonomic resolution would be expected to lead to better explanatory power of the model and stronger inference regarding ecological preferences, even though this could not be tested in the present study. In contrast, for the taxa Cnidaria, Hymenoptera, and Lepidoptera, which have few presence data points in the biomonitoring data, finer taxonomic resolution would likely not lead to a higher information content in the data.

Further research directions

Further interdisciplinary collaboration, particularly between scientists developing databases and those developing models, will be beneficial for both model and database development. We propose that for some species displaying large prior-to-posterior shifts, particularly the list of 29 taxa identified in our study (Table S3.2), current ecological knowledge in databases be revised. Prior-to-posterior shifts, however, could be caused by diverse reasons

that might require multidisciplinary attention to investigate. Reasons for prior-to-posterior shifts include deviations in interpretation of affinity scores in the databases and their conversion to ecological preference scores for modeling, biased estimates of environmental conditions (which often need to be modeled themselves), and regional differences in ecological preferences for the same taxa (e.g., due to adaptation or sub-species with different preferences). Moreover, discrepancies in the predicted and observed distribution of taxa may relate to observational errors in biomonitoring data and additional confounding environmental or anthropogenic factors or processes (e.g., dispersal limitations or historic and random events; Chave 2004, Van de Meutter et al. 2007, Kozak et al. 2008, HilleRisLambers et al. 2012, Cadotte and Tucker 2017) that are currently not included in the model. The modeled distribution of *N. minima* and *N. mortoni* across Switzerland (Fig. 7A, B), for example, showed some false positive locations where our model predicted presence but none was observed. An investigation of what sets these specific locations apart could offer insight to additional factors influencing occurrence, which provides an example of the potential synergy between modeling-based and field-based ecology.

Further development of HS-MSDMs will benefit from increased alignment in the way affinity scores are presented across databases and how they are mathematically formulated in models. Continued input and efforts by taxonomists and ecologists to extend current ecological preference databases will also provide more complete data and support more accurate modeling. For example, an extended temperature database (Halle et al. 2016) could improve model performance by preventing the need to derive temperature preference data for some individual taxa from taxonomically related taxa, as was done in this study. Additionally, in our study we selected 1 ecological database for each specific ecological preference, but an alternative strategy would be to merge prior information from multiple ecological databases to maximize the

available prior information. Preference information is often delineated based on a process of literature review and expert opinion (Schmidt-Kloiber and Hering 2015, Serra et al. 2016), which pools information from a large range of sources and geographic areas. Also, ecological preference data is often represented categorically, which facilitates comparisons across taxa. However, definitions of environmental conditions delineating class boundaries may sometimes be fuzzy. Moreover, the way ecological preferences are derived from ecological databases, normalized, and transformed into a habitat suitability function, like in the HS-MSDM modeling process, leaves room for increased uncertainties in model outputs. For example, the temperature preference scores within the database for *Centroptilum luteolum* were 0 for very cold, 1 for cold, 3 for moderate, 2 for warm, and 4 for eurytherm (where the eurytherm class does not correspond to a specific temperature interval but, rather, represents the ability of a taxon to occur at a wide range of temperatures). However, because the eurytherm temperature class obtained the maximum score of 4, this resulted—after normalization and integration into a habitat suitability function (which integrated the eurytherm class into the scores of the other classes; Appendix S1.3, S1.4)—in a prior distribution that predicted high suitability across all temperature classes. The posterior distribution, showing a lowered suitability for very cold and cold classes, matched more closely with the original scores given in the freshwaterecology.info database for these classes, which illustrates the challenge of interpreting and integrating preference information into a mathematical framework.

A good coverage of all environmental factors, including coverage of different combinations of environmental factors that often naturally co-occur, is difficult to achieve in biomonitoring data unless it is accounted for in the site-selection strategy. Saprobic conditions, for instance, often reflect an elevational gradient. Rivers of higher elevation tend to belong to the

oligosaprobic class and those of lower elevation to the beta-mesosaprobic class, which is largely because saprobity naturally increases from source to mouth. Even though we avoided including strongly correlated environmental factors in our model, such naturally occurring patterns are hard to avoid. For example, sampling in the BDM biomonitoring program is organized along a regular grid spread across Switzerland, independent of patterns in environmental factors. The identification and inclusion of additional sites that lead to a better coverage of all combinations among environmental factors, especially the inclusion of sites with low temperature and higher saprobic values or higher insecticide pollution, would help resolve potentially confounding effects.

Another challenge in building species distribution models for biomonitoring data is variable selection, which needs to include consideration of uncertainty and multi-collinearity in environmental factors. The 6 environmental variables considered in this study are known to affect macroinvertebrate distributions and to vary across Switzerland, including across the Swiss Plateau and the Alps. Inclusion of additional factors in the model, such as hydrological regime and availability of different food sources, could add additional explanatory power, yet their inclusion needs careful consideration. Factors that require estimating, such as the 4 environmental factors we estimated for this study, add uncertainty to model results. For example, stream water temperature was modeled from indirect variables, which themselves include uncertainty. The application of improved temperature measuring systems (e.g., low-cost, remote, environmental sensor platforms; Lockridge et al. 2016) or models would increase the reliability of temperature predictions in future modeling efforts. For environmental factors with partly overlapping information content (e.g., substratum and morphological assessment), the inclusion of both factors influences the strength of the response to each factor compared with a model that

includes only 1 of them. Modeling efforts should consider such multi-collinearity during variable selection by, for example, considering statistical techniques for variable selection and the ecological importance of each factor.

Our model-based approach to revise and complement ecological preference information by confronting it with independent biomonitoring data can be transferred to other organism groups for which ecological preference information and biomonitoring data exist. Our results show that the model works best for taxa with an intermediate frequency of occurrence, roughly between 0.2 and 0.8. Rare taxa can still be included in the model, but a high explanatory power of the model cannot be expected for rare taxa, and, therefore, model outputs will not reveal as much about their ecological preferences.

ACKNOWLEDGEMENTS

Author contributions: PV, PR, and NS contributed to conception and design, data processing, and analysis. WG, PL, and ASK especially contributed to data interpretation, advice, and discussion regarding ecological preferences and macroinvertebrate fauna. All authors contributed to interpretation and drafting the article. PV, NS, WG, PL, and ASK revised the article. All data sources are provided in the text. The monitoring data can be obtained from the MIDAT database (Center for Swiss Cartography of Fauna [CSCF], Neuenburg, Switzerland), and ecological preference data can be obtained from www.freshwaterecology.info.

This study was part of the AQUACROSS project, funded by the European Union's Horizon 2020 research and innovation program (grant agreement No. 642317). We thank the Swiss Federal Office of the Environment (especially Yael Schindler), the Swiss Biodiversity Monitoring Switzerland coordination office, Hintermann and Weber, and the CSCF (especially Nadine Remund) for access to data and support. We thank Rosi Siber, Raoul Schaffner, Ruth Scheidegger, Karin Ghilardi, Bogdan Caradima, and Mathias Kuemmerlen for collaboration and data preparation and Simon Spycher, Christian Stamm, and Irene Wittmer for insights into pesticide pollution.

LITERATURE CITED

Arscott, D. B., J. K. Jackson, and E. B. Kratzer. 2006. Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. Freshwater Science 25:977–997.

Beketov, M. A., K. Foit, R. B. Schäfer, C. A. Schriever, A. Sacchi, E. Capri, J. Biggs, C. Wells, and M. Liess. 2009. Spear indicates pesticide effects in streams—Comparative use of species- and family-level biomonitoring data. Environmental Pollution 157:1841–1848.

Bruggeman, J. 2011. A phylogenetic approach to the estimation of phytoplankton traits. Journal of Phycology 47:52–65.

Cadotte, M. W., and C. M. Tucker. 2017. Should environmental filtering be abandoned? Trends in Ecology and Evolution 32:429–437.

Cao, Y., D. P. Larsen, and R. St-J. Thorne. 2001. Rare species in multivariate analysis for bioassessment: Some considerations. Journal of the North American Benthological Society 20:144–153.

Chave, J. 2004. Neutral theory and community ecology. Ecology Letters 7:241–253.

Cowan, W. 1956. Estimating hydraulic roughness coefficients. Agricultural Engineering 37:473–475.

Dolédec, S., J. M. Olivier, and B. Statzner. 2000. Accurate description of the abundance of taxa and their biological traits in stream invertebrate communities: Effects of taxonomic and spatial resolution. Archives in Hydrobiology 148:25–43.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Graf, W., J. Waringer, and S. U. Pauls. 2009. A new feeding group within larval Drusinae (Trichoptera: Limnephilidae): The *Drusus alpinus*-group sensu Schmid, 1956, including larval descriptions of *Drusus franzi* Schmid, 1956 and *Drusus alpinus* (Meyer-Dür, 1875). Zootaxa 2031:53–62.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, L. Brotons, E. McDonald-Madden, C. Mantyka-Pringle, T. G. Martin, J. R. Rhodes, R. Maggini, S. A. Setterfield, J. Elith, M. W. Schwartz, B. A. Wintle, O. Broennimann, M. Austin, S. Ferrier, M. R. Kearney, H. P. Possingham, and Y. M. Buckley. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424–1435.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Halle, M., A. Müller, and A. Sundermann. 2016. Ableitung von Temperaturpräferenzen des Makrozoobenthos für die Entwicklung eines Verfahrens zur Indikation biozönotischer Wirkungen des Klimawandels in Fließgewässern. KLIWA-Berichte, Heft 20. Arbeitskreis Klimaveränderung und Wasserwirtschaft, LUBW Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg, Bayerisches Landesamt für Umwelt, Landesamt für UmweltRheinland-Pfalz, Deutscher Wetterdienst. (Available from: https://www.kliwa.de/_download/KLIWAHeft20.pdf)

Hering, J. G., E. Hoehn, A. Klinke, M. Maurer, A. Peter, P. Reichert, C. Robinson, K. Schirmer, M. Schirmer, C. Stamm, and B. Wehrli. 2012. Moving targets, long-lived infrastructure, and increasing needs for integration and adaptation in water management: An illustration from Switzerland. Environmental Science & Technology 46:112–118.

HilleRisLambers, J., P. B. Adler, W. S. Harpole, J. M. Levine, and M. M. Mayfield. 2012. Rethinking community assembly through the lens of coexistence theory. Annual Review of Ecology, Evolution, and Systematics 43:227–248.

Höss, S., P. Heininger, E. Claus, C. Möhlenkamp, M. Brinke, and W. Traunspurger. 2017. Validating the NemaSPEAR[%]-index for assessing sediment quality regarding chemical-induced effects on benthic communities in rivers. Ecological Indicators 73:52–60.

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography 21:498–507.

Kefford, B. J., P. K. Botwe, A. J. Brooks, S. Kunz, R. Marchant, S. Maxwell, L. Metzeling, R. B. Schäfer, and R. M. Thompson. 2020. An integrated database of stream macroinvertebrate traits for Australia: Concept and application. Ecological Indicators 114:106280.

Kissling, W. D., L. Dalby, C. Fløjgaard, J. Lenoir, B. Sandel, C. Sandom, K. Trøjelsgaard, and J. Svenning. 2014. Establishing macroecological trait datasets: Digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution 4:2913–2930.

Kozak, K. H., C. H. Graham, and J. J. Wiens. 2008. Integrating GIS-based environmental data into evolutionary biology. Trends in Ecology and Evolution 23:141–148.

Langhans, S. D., J. Lienert, N. Schuwirth, and P. Reichert. 2013. How to make river assessments comparable: A demonstration for hydromorphology. Ecological Indicators 32:264–275.

Lenat, D. R., and V. H. Resh. 2001. Taxonomy and stream ecology—The benefits of genus- and species-level identifications. Journal of the North American Benthological Society 20:287–298.

Liechti, P. 2010. Methoden zur Untersuchung und Beurteilung der Fliessgewässer: Chemisch-physikalische Erhebungen, Nährstoffe. Umwelt-Vollzug. Bundesamt für Umwelt, Bern, Switzerland. (Available from: https://www.modul-stufen-konzept.ch/download/ChemieD_Juni2010.pdf)

Liess, M., R. B. Schäfer, and C. A. Schriever. 2008. The footprint of pesticide stress in communities—Species traits reveal community effects of toxicants. Science of the Total Environment 406:484–490.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2007. AUC: A misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17:145–151.

Lockridge, G., B. Dzwonkowski, R. Nelson, and S. Powers. 2016. Development of a low-cost Arduino-based sonde for coastal applications. Sensors 16:528.

Losos, J. B. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecology Letters 11:995–1007.

Menezes, S., D. J. Baird, and A. M. V. M. Soares. 2010. Beyond taxonomy: A review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. Journal of Applied Ecology 47:711–719.

Nijboer, R. C., and A. Schmidt-Kloiber. 2004. The effect of excluding taxa with low abundances or taxa with small distribution ranges on ecological assessment. Hydrobiologia 516:347–363.

Poff, N. L., J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: Trait-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25:730–755.

Robinson, C. T., N. Schuwirth, S. Baumgartner, and C. Stamm. 2014. Spatial relationships between land-use, habitat, water quality and lotic macroinvertebrates in two Swiss catchments. Aquatic Sciences 76:375–392.

Ruaro, R., E. A. Gubiani, A. M. Cunico, Y. Moretto, and P. A. Piana. 2016. Comparison of fish and macroinvertebrates as bioindicators of Neotropical streams. Environmental Monitoring and Assessment 188:45.

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

Schäfer, R. B., B. J. Kefford, L. Metzeling, M. Liess, S. Burgert, R. Marchant, V. Pettigrove, P. Goonan, and D. Nugegoda. 2011. A trait database of stream invertebrates for the ecological risk assessment of single and combined effects of salinity and pesticides in South-East Australia. Science of the Total Environment 409:2055–2063.

Schmidt-Kloiber, A., and D. Hering. 2015. www.freshwaterecology.info—An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. Ecological Indicators 53:271–282.

Schmidt-Kloiber, A., and R. C. Nijboer. 2004. The effect of taxonomic resolution on the assessment of ecological water quality classes. Hydrobiologia 516:269–283.

Serra, S. R. Q., F. Cobo, M. A. S. Garça, S. Dolédec, and M. J. Feio. 2016. Synthesising the trait information of European Chironomidae (Insecta: Diptera): Towards a new database. Ecological Indicators 61:282–292.

Stribling, J. B., B. K. Jessup, and D. L. Feldman. 2008. Precision of benthic macroinvertebrate indicators of stream condition in Montana. Freshwater Science 27:58–67.

Stucki, P. 2010. Methoden zur Untersuchung und Beurteilung der Fliessgewässer. Macrozoobenthos – Stufe F (flächendeckend), Umwelt-Vollzug Nr. 1026. Bundesamt für Umwelt, Bern, Switzerland. (Available from: https://www.modul-stufen-konzept.ch/download/MZB_Stufe_F-D_20111215.pdf)

Tachet, H., P. Richoux, M. Bournaud, and P. Usseglio-Polatera. 2010. Invertébrés d`eau douce systématique, biologie, écologie. CNRS éditions, Paris, France.

Van de Meutter, F., L. De Meester, and R. Stoks. 2007. Metacommunity structure of pond macroinvertebrates: Effects of dispersal mode and generation time. Ecology 88:1687–1695.

Vermeiren, P., P. Reichert, and N. Schuwirth. 2020. Integrating uncertain prior knowledge regarding ecological preferences into multi-species distribution models: Effects of model complexity on predictive performance. Ecological Modelling 420:108956.

Vieira, N. M. K., N. L. Poff, D. Carlisle, S. R. Moulton II, M. K. Koski, and B. C. Kondratieff. 2006. A database of lotic invertebrate traits for North America. United States Geological Survey Data Series 187, Reston, Virginia. (Available from: https://pubs.water.usgs.gov/ds187)

Vivien, R., F. Lejzerowicz, and J. Pawlowski. 2016. Next-generation sequencing of aquatic oligochaetes: Comparison of experimental communities. PLoS ONE 11:e0148644.

FIGURE CAPTIONS

Fig. 1. Derivation of a habitat suitability *h* for temperature (A), flow velocity (B), saprobic condition (C), substratum classes (1: pebbles, 2: gravel, 3: sand and silt, 4: mud, 5: roots and litter, 6: microphytes (algae), 7: macrophytes) (D), insecticide pollution (E), and morphology (F). Panels A–D show the original prior information from the databases, in the form of affinity scores as numbers at the top of each panel, for an example taxon (*Baetis alpinus*), ecological preference scores s_{jr}, normalized to the habitat-suitability interval 0 to 1 (gray continuous lines), and processed to obtain a habitat suitability function h_r (black dashed lines). Panels E and F show habitat suitability as a continuous function of insecticide pollution and morphology, respectively, for an insensitive (continuous line), moderately sensitive (dashed line), and sensitive (dotted line) taxon. Given the environmental conditions, x_{it_i}, at site i and sampling t_i, the habitat suitabilities, h_{it_ijr}, can be derived from the suitability functions.

Fig. 2. Conceptual representation of the HS-MSDM based on habitat suitabilities *h*. Ecological preferences and parameter α are shown in ovals, and environmental factors are shown in hexagons. The β-parameters are weighting factors that apply to the whole assemblage (i.e., are not taxon-specific). The α-parameters, as well as the ecological preference scores (dashed lines), are taxon-specific. Model output is the probability of occurrence for each taxon after logistic transformation (modified after Vermeiren et al. 2020).

Fig. 3. Comparison of prior and posterior probability distributions for ecological preferences (in this example, temperature) for a unique taxon. The bold dashed line indicates the ecological preference score, which was derived (and normalized) from a database. The gray area shows the prior distribution that reflects prior knowledge about the ecological

preference score and its uncertainty. The solid black line illustrates the marginal posterior distribution obtained from Bayesian inference, updating the prior distribution by confrontation with independent data through model calibration. In this example, the prior-to-posterior shift to higher values indicates higher preferences for very cold and moderate temperatures than were suggested by the database, whereas the low preference for warm temperatures is confirmed by a reduction in parameter uncertainty (the posterior distribution is narrower than the prior and has the maximum at the same value as the prior). A posterior distribution that is similar to the prior indicates low information content in the data or a minor influence of the parameter on model output.

Fig. 4. Standardized deviance d_j (A) and D_j^2 statistics (B) for model MS1 fit of individual taxa vs their relative frequency of occurrence for: Ephemeroptera (n = 54), Plecoptera (n = 42), Trichoptera (n = 76) (EPT), and remaining taxa (n = 71) in Swiss rivers. Black lines indicate the null model. Taxa abbreviations: E: Elmidae, G: Gammaridae, Br: *Baetis rhodani*, Ba: *Baetis alpinus*, Rt: *Rhyacophila tristis*, Nm: *Nemoura mortoni*, Pl: *Protonemura lateralis*, Lb: *Leuctra braueri*, and Dd: *Drusus discolor*. The distribution of standardized deviance d_j (C) and D_j^2 statistics (D) are boxes encompassing the 75th and 25th percentiles with the medians as horizontal bars and whiskers extending 1.5× the interquartile range for model MS1 for EPT taxa. Green boxplots include all taxa within the respective order, and orange boxplots include only taxa with a relative frequency of occurrence between 0.2 and 0.8. The numbers on top of the boxplots provide the mean values.

Fig. 5. Distributions of the β_r parameters, which indicate the strength of the effect of each environmental factor, via the ecological preferences, on the probability of occurrence of

macroinvertebrates across Swiss rivers. Priors are shown as yellow shaded areas, which are hardly visible because they are much wider (flatter) than the posterior distributions. Priors are normally distributed with mean and standard deviation of 3.

Fig. 6. Distribution of prior-to-posterior shifts across all taxa (dark gray) and across taxa with a relative frequency of occurrence of 0.2 to 0.8 (light gray) for each class of the different ecological preferences across Swiss rivers: temperature preference (A), flow-velocity preference (B), saprobity (C), sensitivity to pesticides (D), and substratum preference (E). Note: some of these taxa derived information from genus or family level. Those that did not have or did not derive information (i.e., with NA values when entering the model) are not plotted. Gray horizontal lines indicate a prior-to-posterior shift of ± 0.1. Boxes encompass the 75^{th} and 25^{th} percentiles with the medians as horizontal bars and whiskers extending 1.5× the interquartile range.

Fig. 7. Example results of HS-MSDM MS1 for 4 species. Maps show the distribution of the species: *Nemoura minima* (A), *Nemoura mortoni* (B), *Rhyacophila tristis* (C), and *Rhyacophila hirticornis* (D), across Swiss rivers with presence (blue) and absence (red) observations. Point size increases with increased predicted probability of occurrence. Plots below the maps show the ecological preferences of the species for temperature and flow velocity with prior information as red horizontal lines (and parameterized as gray shaded areas) and posterior marginal parameter distribution in black, or colored for the 3 cross-validation runs. Note that some taxa have fewer data points because of taxonomic mismatches in the data (see Methods: Invertebrate biomonitoring data).

Fig. 8. Comparison of the marginal posterior distributions of the habitat suitability parameters ($\boldsymbol{s}_r$) for temperature (T), flow velocity (v), saproby (sap), insecticides, and substratum

(subst) for the family Baetidae (model MF1, black line) with the individual species from the same family (model MS1, colored lines) for species with a relative frequency of occurrence >0.05. The prior distribution of the family is shown as gray shaded area. The prior distributions of the species are given as colored areas if they deviate from the prior distribution of the family. If species-level prior distributions are the same as the family level, they are not shown (i.e., they are the same as the gray shaded area). The x-axis shows the parameter value (between 0–1) and the y-axis shows the probability density (between 0–1). Numbers in the legend indicate the relative frequency of occurrence.

Table 1. Overview of the BDM dataset including all taxa at the most detailed available taxonomic resolution, with most Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa resolved to species or genus level and other taxa to coarser levels (S) and the same data with EPT species and genera merged to family level (F).

Dataset	Finest taxonomic level	Total taxa	Taxa at species level	Taxa at genus level	Taxa at family level	Coarser level taxa	Sites	Samplings
S	Species	245	148	23	68	6	491	579
F	Family	102	–	–	96	6	481	562

Table 2. Environmental factors, their derivation from indirect environmental factors, and the ecological preference to which they were linked. FWE refers to the freshwaterecology.info database (Schmidt-Kloiber and Hering 2015). Additional details are provided in Appendix S1.2 and Vermeiren et al. (2020).

Temperature *T* (°C)
Definition: Maximum morning water temperature in summer as defined by the FWE database.
Derivation: Estimated from catchment area and mean catchment elevation (Vermeiren et al. 2020).
Ecological preference: Temperature preference (FWE database) described in 4 classes after normalization and preprocessing: very cold <6°C, cold 6–10°C, moderate 10–18°C, warm >18°C.
Saprobic condition *sap* (unitless)
Definition: Water quality related to easily degradable organic substances leading to reduced oxygen conditions for macroinvertebrates.
Derivation: Estimated from proportion of agricultural land in the catchment, fraction of treated wastewater, and livestock unit densities (Vermeiren et al. 2020).
Ecological preference: Saprobity (FWE database, using the Austrian saprobity values). Sensitivity of organisms to pollution by easily degradable organic substances is described by saprobity, which contains 5 classes from no to high pollution (xeno-saprobic = 0, oligo-saprobic = 1, β-meso-saprobic = 2, α-meso-saprobic = 3, and poly-saprobic = 4).
Flow velocity *v* (m/s)
Definition: Average flow velocity of the river reach.

Derivation: Estimated from slope, mean annual discharge, river width at the sampling location, width variability, the proportion of the substratum containing sediments with grain size >2.5 mm, and the proportion of the riverbed containing macro-algae.

Ecological preference: Flow velocity preference (Tachet database, there called current velocity preference) described in 4 classes: standing <0.05 m/s, low 0.05–0.24 m/s, moderate 0.25–0.5 m/s, and high >0.5 m/s.

Insecticide pollution *IP* (unitless)

Definition: A combination of agricultural and urban sources that contribute to insecticide pollution in the river.

Derivation: The insecticide application rate (IAR) is a weighted sum of the proportions of crops weighted by the average number of insecticide applications/year; the wastewater fraction (WW) is the proportion of wastewater at average discharge conditions. Insecticide pollution is calculated as the weighted sum of IAR and WW, considering a weighting factor that was estimated from the data (see Table S1.2).

Ecological preference: Sensitivity regarding pesticides (Spear database), which is a binary classification (sensitive or insensitive) from a combination of 4 biological traits that influence the sensitivity of macroinvertebrates to pesticide (mainly insecticide) pollution (sensitivity to organic chemicals, generation time, migration ability, and presence of aquatic life stages during the application period of insecticides; Liess et al. 2008). Note: in the main text, referred to as sensitivity to insecticide pollution.

Substratum classes *subst* [0:1]

Definition: Relative coverage of substratum types.

Derivation: % cover of each substratum class recorded in the field including: 1) pebbles (containing mobile blocks >250 mm, natural and artificial surfaces, and coarse inorganic sediments between 25 mm and 250 mm), 2) macrophytes (containing: mosses, hydrophytes, and helophytes), 3) gravel between 2.5 mm and 25 mm, 4) roots and litter, 5) sand and silt <2.5 mm, 6) mud <0.1 mm, and 7) microphytes (algae).

Ecological preference: Microhabitat/substratum preference (Tachet database), which was available in 9 discrete classes. These were combined into 7 classes to link them to substratum classes recorded in the field. Specifically, we combined the classes sand and silt, and roots and litter, respectively, based on the mean of the normalized affinity score across the 2 classes.

Morphology *morph* [0:1]

Definition: Morphological assessment according to the Swiss modular concept for stream assessment including width variability, modifications of the river bed and the river banks, width and condition of the riparian zone, and presence of culverts. 0 is the worst and 1 the best morphological condition (Liechti 2010, Langhans et al. 2013).

Ecological preference: Not available.

Table 3. Performance of different HS-MSDM applications for fitting and predicting spatial-distribution patterns of macroinvertebrate assemblages at river sites throughout Switzerland, given different inputs of BDM biomonitoring and environmental data for the 2 datasets S and F (see Table 1). Total number of data points (*n* dat), number of estimated parameters (*n* par), mean deviance for model calibration (*d* fit) and cross-validation (*d* val ± SD across the 3 folds), explanatory power for calibration (D^2 fit) and cross-validation (D^2 val ± SD across the 3 folds), and the AUC for cross-validation (AUC val ± SD across the 3 folds). T = temperature, v = flow velocity, sap = saprobic condition, IP = insecticide pollution, subst = substratum classes, and morph = morphology.

Model version	Finest taxonomic level	Environmental factors	*n* dat	*n* par	*d* fit	*d* val	D^2 fit	D^2 val	AUC val
MS1	Species	T, v, sap, IP, subst	111,419	5353	0.360	0.384 ± 0.009	0.183	0.177 ± 0.006	0.760 ± 0.015
MS2	Species	T, v, sap, IP, morph	115,652	3927	0.359	0.380 ± 0.008	0.178	0.173 ± 0.011	0.796 ± 0.006
MS3	Species	T, v, sap, IP, subst, morph	111,419	5597	0.364	0.386 ± 0.004	0.176	0.172 ± 0.007	0.771 ± 0.014
MF1	Family	T, v, sap, IP, subst	57,324	2251	0.463	0.489 ± 0.009	0.148	0.143 ± 0.003	0.733 ± 0.010

Table 4. Overview of the completeness of data on ecological preferences, including the relative frequency of occurrence of taxa with missing affinity scores and the taxonomic level at which their ecological preferences were derived, which were included in the HS-MSDM MS1 model.

Ecological preference	Taxa with missing affinity scores	Mean relative frequency of occurrence of taxa with missing affinity scores (±SD)	Taxa with prior information derived from genus level	Taxa with prior information derived from family level
Temperature	82	0.14 ± 0.22	34	28
Flow velocity	11	0.16 ± 0.28	131	6
Saprobity	37	0.14 ± 0.21	131	7
Insecticide pollution	11	0.16 ± 0.28	73	21
Substratum	11	0.16 ± 0.28	131	6

Table 5. Classification of taxa into sensitive and insensitive categories for morphology (mean posterior ecological preference score for morphology >0.55 and <0.45, respectively, in models MS2 and MS3). We excluded taxa with a relative frequency of occurrence <0.1 and a D^2 <0.2. Numbers following species names indicate their relative frequency of occurrence in the full dataset (MS2). Taxa with * obtained a relative frequency of occurrence <0.1 and a D^2 <0.2 in model MS2 but not in MS3.

Taxa sensitive to morphology	Taxa insensitive to morphology
Baetis alpinus, 0.74	*Amphinemura*, 0.22
Baetis muticus, 0.47	*Baetis lutheri*, 0.12
Blephariceridae, 0.27	*Ecdyonurus venosus*, 0.17
Capnioneura nemuroides, 0.11*	Elmidae, 0.53
Chloroperla susemicheli, 0.13*	Gammaridae, 0.45
Drusus discolor, 0.50	*Hydropsyche siltalai*, 0.12
Ecdyonurus helveticus, 0.31	*Paraleptophlebia submarginata*, 0.12
Epeorus alpicola, 0.16	*Protonemura intricata*, 0.14
Epeorus assimilis, 0.16	*Rhyacophila tristis*, 0.55
Habroleptoides confusa, 0.27*	Tinodes, 0.21
Leuctra braueri, 0.43	
Leuctra major, 0.17	
Leuctra nigra, 0.28	
Nemoura minima, 0.31	
Nemoura mortoni, 0.50	
Protonemura lateralis, 0.53	

Protonemura nimborum, 0.13
Rhithrogena loyolaea, 0.23*

Supporting information:

Confronting existing knowledge on ecological preferences of stream macroinvertebrates with independent biomonitoring data using a Bayesian multi-species distribution model

Peter Vermeiren[1,3], Peter Reichert[1], Wolfram Graf[2], Patrick Leitner[2], Astrid Schmidt-Kloiber[2], Nele Schuwirth[1]

[1]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

[2]BOKU: University of Natural Resources and Life Sciences, Institute of Hydrobiology and Aquatic Ecosystem Management, Gregor-Mendel-Straße 33, 1180 Vienna, Austria

[3]Present address: Radboud University, Dept. of Environmental Science, Heyendaalseweg 135, 6525HP Nijmegen, The Netherlands

E-mail: peter.vermeiren@gmail.com

Appendix S1. Methodological information

Appendix S2. HS-MSDM model performance and summary statistics

Appendix S3. Review of ecological preferences

Appendix S1: Methodological information

S1.1. Macroinvertebrate sampling procedure

The sampling followed the Swiss method for stream assessment (Stucki 2010). Wadeable river reaches with a Strahler order ≥2 that appear on maps with a 1:25,000 scale are sampled at a length of 10× the average width of the river section between the epirhithral and epipotamal zones (reaches are, on average, 2.9 m wide). Eight kick samples are collected in separate microhabitats/reach in an effort to represent all habitat types within the sampling site. Sampling is conducted from mid-February to mid-August in periods with low hydrological disturbance and normal weather conditions following an altitude-stratified sampling window containing 5 strata of 400 m starting at 200 m. Sampling is conducted within a 1-month core period with half month buffers before and after. Macroinvertebrates collected are pooled per reach and maintained in alcohol until identification in the laboratory.

S1.2 Links between ecological preferences and environmental factors

The ecological preferences and their linked environmental factors used in this study are presented in Table S1.2. We chose to use environmental factors that were expected to have a direct influence on organisms and, thus, a direct link to their ecological preferences. The use of direct influence factors facilitates the establishment of cause-effect relationships and, thus, has the potential to result in more universal models compared to those based on indirect influence factors (Guisan and Zimmermann 2000). Nonetheless, we needed to derive direct environmental factors from indirect ones in cases where measurements of direct environmental factors were absent.

Additional direct environmental factors were considered but not included in the final model. Specifically: the proportion of artificial bed modification was considered, but its effects were not well captured in the invertebrate datasets. Indeed, samples taken with the multi-microhabitat sampling method do not reflect the relative proportion of artificial habitats (e.g., due to bed modification) in the river. The condition of the riparian zone was considered but did not reveal a significant effect in preliminary analyses and did not cover a large environmental range in our sampling sites. Hydropeaking data and data related to water abstraction were considered, but only coarse data were available, and few sites were affected. For these reasons, these environmental factors were not considered in the final model.

Table S1.2. Environmental factors, their derivation from indirect influence factors, and the ecological preference to which they were linked. FWE refers to the freshwaterecology.info database (Schmidt-Kloiber and Hering 2015).

Temperature *T* (°C)

Definition: Maximum morning water temperature in summer as defined by the FWE database.

Derivation: Derived using a linear model, which was calibrated with hourly water temperature measurements from 58 stations between 2006 and 2015 across Switzerland (Appendix S1 in Vermeiren et al. 2020).

$$T_{\text{max morn.}} = 18.4°C + 2.77°C \cdot \log_{10}(\text{catchment area}/1\text{km}^2) - 0.006\frac{°C}{m} \cdot \text{mean catchment elevation}$$

Ecological preference: Temperature preference (FWE database) described in 4 continuous classes after normalization and preprocessing: very cold <6°C, cold 6–10°C, moderate 10–18°C, warm >18°C; the eurytherm class was integrated into the other classes after normalization of the ecological preference score.

Saprobic condition *sap* (-)

Definition: Water quality related to easily degradable organic substances leading to reduced oxygen conditions for macroinvertebrates.

Derivation: The Saprobic Index according to Zelinka and Marvan (1961) is a standard for water quality assessment based on biological–ecological indicators (Bernatowicz et al. 2009). We derived a saprobic condition on a continuous scale from 0 to 4 that is based on environmental factors in the catchment (Appendix S1 in Vermeiren et al. 2020). We estimated the water quality corresponding to the Austrian saprobic system (hereafter: Saprobic condition) using species classifications from the Fauna Aquatica Austriaca (Moog 2002) based on a linear model using indirect influence factors that would affect the amount of organic matter, such as the proportion of agricultural land in the catchment, the wastewater fraction, and the life-stock density. We developed this by linking O_2-diss, NH_4^+, NO_3^-, PO_4^{3-}, and DOC measurements to environmental predictors at 345 stations between 2009 and 2015 across Switzerland. Subsequently, we used this model to predict the saprobic condition for the macroinvertebrate monitoring sites in our datasets.

$$\text{Saprobic condition} = 0.746 + 0.0182 \cdot p_{\text{agri}} + 4.427 \cdot \text{WW} + 0.00668 \cdot \text{LUD}$$

with p_{agri} proportion of agricultural land in the catchment, WW fraction of treated waste water, and LUD livestock unit densities.

Ecological preference: Saprobity (FWE database, using the Austrian saprobity values). The sensitivity of organisms to pollution by easily degradable organic substances is described by the saprobity, which contains 5 continuous classes from no to high pollution (xeno- = 0, oligo- = 1, β-meso- = 2, α-meso- = 3 and poly-saprobic = 4).

Flow velocity *v* (m/s)

Definition: Average flow velocity of the river reach.

Derivation: The flow velocity v was estimated using the following equation (Chow 1959):

$$v = \left(\sqrt{S}/n\right)^{\frac{3}{5}} \cdot (Q/W)^{\frac{2}{5}}$$

where S is the slope (%, obtained from a large-scale topographic landscape model for Switzerland, Swisstopo 2016), Q is the mean annual discharge (m/s, obtained from a runoff and flow regime dataset FOEN 2013) and W is the river width (m, measured at the sampling location). The Manning's

coefficient, n (s/m$^{1/3}$), was set to 0.08 s/m$^{1/3}$ or estimated when sufficient data were available as follows (based on estimates by Cowan 1956):

$$n = (0.03 + 0.03 \cdot W\text{var} + 0.03 \cdot \text{Sed} + 0.05 \cdot \text{Alg}) \text{ s/m}^{1/3}$$

where Wvar is the width variability (between 0 (low) and 1 (high)), Sed is the proportion of the substratum containing sediments with grain size >2.5 mm, and Alg is the proportion of the riverbed containing macro-algae (between 0 (low) and 1 (high)).

Ecological preference: Current velocity preference (Tachet database) described in 4 continuous classes: standing <0.01 m/s, low <0.25 m/s, moderate 0.25–0.5 m/s, and high >0.5 m/s.

Insecticide pollution *IP* (-)

Definition: A combination of agricultural and urban sources of insecticides that contribute to insecticide pollution in the river. Specifically: Insecticide Application Rate (IAR) and the Wastewater fraction (WW).

Derivation: The IAR is a weighted sum of the proportions of crops:

$$IAR = \sum_{i=1}^{n} p_{\text{crop } i} \cdot \text{a}_i$$

with $p_{\text{crop } i}$ proportion of land-use of crop i in the catchment and a_i average yearly number of insecticide applications for crop i which are 1.83 for rapeseeds, 2.66 for vegetables, 3.10 for orchards, 0.37 for vineyards, 0.44 for potatoes, 0.03 for cereals, 0.38 for legumes, 0.07 for beets, and 0.01 for corn. Land-use data were obtained from the Swiss Federal Statistical Office (BFS 2008a, b).

The wastewater fraction, WW, is given as the proportion of wastewater (m^3/s) per mean annual discharge (m^3/s). Data were obtained from a runoff and flow regime dataset FOEN (2013).

IAR and UA were combined using a weighting factor (U_{IAR}) reflecting equal contributions of each source (Wittmer et al. 2010). The value of the weighting factor was set to 0.6 corresponding to the average weighting factor of different agricultural crops given their proportional area in the BDM dataset.

Ecological preference: Sensitivity regarding pesticides (Spear database), which is a binary classification (sensitive or insensitive) from a combination of 4 biological traits that influence the sensitivity of macroinvertebrates to pesticide (mainly insecticide) pollution (sensitivity to organic chemicals, generation time, migration ability, and presence of aquatic life stages during the application period of insecticides; Liess et al. 2008). Note: in the main text, referred to as sensitivity to insecticide pollution.

Substratum classes *subst* [0:1]

Definition: Relative coverage of substratum types.

Derivation: Percentage cover of each substratum class recorded in the field including: 1) pebbles (containing mobile blocks >250 mm, natural and artificial surfaces, and coarse inorganic sediments between 250 mm and 25 mm), 2) macrophytes (containing: mosses, hydrophytes and helophytes), 3) gravel between 25 mm and 2.5 mm, 4) roots and litter, 5) sand and silt <2.5 mm, 6) mud <0.1 mm, and 7) microphytes (algae).

Ecological preference: Microhabitat/substratum preference (Tachet database) which was available in 9 discrete classes. These were combined into 7 classes in order to link to substratum classes recorded in the field. Specifically, we combined the classes sand and silt, and roots and litter, respectively, based on the mean of the normalized affinity score across the 2 classes.

Morphology *morph* [0:1]

Definition: Morphological assessment according to the Swiss modular concept for stream assessment, including width variability, modifications of the river bed and the river banks, width and condition of the riparian zone, and presence of culverts; 0 is the worst and 1 the best morphological condition (Liechti 2010, Langhans et al. 2013).

Ecological preference: Not available.

S1.3. Ecological preference normalization

After extracting the affinity scores from the databases, we normalized all scores between 0 and 1 by dividing by the maximum score across all classes for the given ecological preference. For example, preference for temperature was described in 5 classes: "very cold" (*vc*): <6°C; "cold" (*c*): 6–10°C; "moderate" (*m*): 10–18°C; "warm" (*w*): >18°C; and "eurytherm" (*eu*). (In this context, the ecological preference for temperature presented a unique case because it included a eurytherm class, which does not correspond to a specific temperature interval but, rather, represents the ability of a taxon to occur at a wide range of temperatures. None of the other ecological preferences in the current study included such an "indifferent" class). For each taxon, 10 points were distributed over these 5 classes. To derive normalized ecological preference s_{jT} between 0 and 1 for each taxon, j, regarding temperature, T, the maximum points, P_{jT}, between each of the 4 temperature-specific classes and the indifferent class was divided by the maximum number of points over all classes, as illustrated in equation S1.3 for the case of the "very cold" class and in table S1.3.

$$s_{jT}^{vc} = \frac{\max\left(P_{jT}^{vc}, P_{jT}^{eu}\right)}{\max\left(P_{jT}^{vc}, P_{jT}^{c}, P_{jT}^{m}, P_{jT}^{w}, P_{jT}^{eu}\right)} \quad \text{(Eq. S1.3)}$$

Table S1.3. Example of ecological preference normalization for the temperature preference

	Original affinity scores					**Ecological preferences after normalization**			
	Very cold	Cold	Moderate	Warm	Eurytherm	Very cold	Cold	Moderate	Warm
Taxon A	10	0	0	0	0	1	0	0	0
Taxon B	2	2	0	0	6	1	1	1	1
Taxon C	6	0	0	2	2	1	0.3	0.3	0.3

For other ecological preferences, r, the normalized ecological preference, s_{jr}, between 0 and 1, was derived by dividing the affinity scores in each class by the maximum number of points over all classes.

S1.4. Formulation of habitat suitability functions

To integrate ecological preferences into species distribution models, we used the concept of habitat suitability as described in the main text (Eq. 1). Depending on the way the ecological preference was available, additional processing steps were required to turn ecological preferences (normalized between 0 and 1) into a habitat suitability.

S1.4.1. Ecological preferences containing multiple continuous classes

For ecological preferences containing multiple continuous classes (in our study: temperature preference, current preference, and saprobity), we linearly interpolated across the preference classes (Fig. 1A–C). This resulted in a suitability profile for each taxon and ecological preference that avoided sharp changes at class boundaries. We chose interpolation points close to the class boundaries to minimize deviance from the original information.

Note: we considered morphology to be a special case where only 1 preference class was available and all values were missing. We chose this formulation because it allowed us to propose a new ecological preference with scores for different taxa, based on the observational data. An alternative approach would have been to enter morphology directly as an influence factor (rather than after transformation into a habitat suitability).

S1.4.2. Ecological preferences containing 1 binary class describing restrictions

For ecological preferences where only 1 binary class was used (in our study: sensitivity to insecticides), we used a function inspired by the Holling type 3 functional response (Holling 1959, Eq. S1.4.2a). In this equation, the suitability profile decreases along an s-curve with increasing values of the corresponding restricting environmental factor (Fig. 1E).

$$h_r(x_{ir}, s_{jr}) = \frac{{K_{jr}}^2}{{K_{jr}}^2 + {x_{ir}}^2} = \frac{1}{1 + {K_{\text{inv},jr}}^2 \cdot {x_{ir}}^2} \quad \text{(Eq. S1.4.2a)}$$

K_{jr} was used to characterize the dependence of the habitat suitability function h_r on the restricting environmental factor x_{ir} corresponding to ecological preference r at site i. K_{jr} specifies the level of the environmental factor at which the habitat suitability was reduced by half. K_{jr} was derived from a universal minimum value for the most sensitive taxa, $K_{\text{min},r}$, divided by the taxon-specific sensitivity information, s_{jr} (Eq. S1.4.2b). Using equation S1.4.2a, we extended the values of species sensitivities from the 2 discrete values of 0 and 1 to a continuous scale of sensitivities between 0 and 1. This procedure resulted in 1 parameter/taxon (s_{jr}) and 1 overall community parameter/ecological preference ($K_{\text{min},r}$). To avoid infinite values of K_{jr} for (completely) insensitive taxa, we parameterized the equations with $K_{\text{inv}} = 1/K$ and $K_{\text{invmax}} = 1/K_{\text{min}}$ (see Eqs S1.4.2a and c).

$$K_{jr} = \frac{K_{\text{min},r}}{s_{jr}} \quad \text{(Eq. S1.4.2b)}$$

$$K_{\text{inv},jr} = K_{\text{invmax},r} s_{jr} \quad \text{(Eq. S1.4.2c)}$$

S1.4.3. Ecological preferences containing multiple discrete classes

For ecological preferences containing multiple discrete classes (in our study: substratum preferences), each class, c, was described by a taxon-specific parameter (normalized score), s_{jr_c}, for that class (Eq. S1.4.3). The habitat suitability function for this trait, r, was then determined by a weighted average of the normalized trait scores for the different classes, s_{jr_c}, multiplied by the proportion of each class, c, within the site x_{ic}

$$h_r(x_{ir}, s_{jr}) = \sum_c s_{jr_c} \cdot x_{ir_c} \quad \text{(Eq. S1.4.3)}$$

(see Fig. 1D).

S1.5. Prior distributions of model parameters

The HS-MSDM used in this paper requires the estimation of a large number of parameters $\theta = (\{\alpha_j\}, \{\beta_r\}, \{u_r\}, \{s_{jr}\})$ (Table 3 main text). However, this does not lead to identifiability problems, as we have good prior knowledge about the ecological preference parameters from the databases and do not want to leave the model the freedom to deviate strongly from these. We chose normal distributions truncated to the interval [0, 1] with a mean centered at the normalized ecological preference from the ecological database and with a standard deviation of 0.2. These informative priors resulted in the use of ecological preferences similar to those in the database except in cases in which there is strong conflicting evidence in the data. Note that for taxa for which no affinity scores could be derived from the databases, a uniform distribution was used as a vague prior for their ecological preference.

For all parameters (except those defining the prior knowledge on ecological preferences), we chose wide priors to primarily learn from the data. Nevertheless, the model formulation given by the equations 2 and 3 in the main text, together with the range of the habitat suitabilities from 0 to 1, indicate reasonable ranges of the parameters of the HS-MSDM of invertebrate occurrences. We calculated a range of z-values Δz_{90} for which the probability increases from 5% to 95%. The variation in $\beta_r\ h_{it_ijr}$ for each trait and across all sites should 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 $\Delta z_{90}/2$. As we expect a positive response of the probability of occurrence with increasing habitat suitability, we assume positive prior means of $\Delta z_{90}/2$. To avoid "inverse interpretation" of the habitat suitabilities compensated by negative values of the parameters β_r, we truncated the priors of the parameters β_r at zero. We considered 5 to 6 habitat suitability functions and the parameter α_j must be able to adjust the average probability. As the habitat suitabilities 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 α_j with a 5 (or 6) times larger, but negative mean, $-5\Delta z_{90}/2$, and used half of the absolute mean as the standard deviation.

In the model application, we included 2 additional parameters u_r to model the response to insecticides. We estimated a mean weighting factor to combine the effect of the proportion of wastewater fractions (WW, urban insecticide source) with insecticide application rates (IAR, agricultural insecticide source) to be 0.6 (see also Appendix S1.2 on how to combine IAR and WW). We therefore chose a lognormal prior with mean and standard deviation equal to 0.6 for this weighting factor. The parameter K_{min} 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 50$^{\text{th}}$ percentile of the data). As we were uncertain about its value, we used a lognormal distribution for K_{invmax} with mean and standard deviation equal to 10 (see Appendix S1.2 for the meaning of this parameter). Note that we parameterized the lognormal distribution here with its mean and standard deviation and not, as it is done in many software packages, with the mean and standard deviation of the log of the parameter.

The joint prior was formulated by assuming independence. This is usually done when no prior information about the correlation structure of the individual parameters is available.

S1.6. Numerical solution

Bayesian inference for the HS-MSDMs was conducted using Hamiltonian Monte Carlo Markov Chain sampling (Duane et al. 1987, Brooks et al. 2011) using 4 chains, each with 12,000 iterations, including a warm-up phase of 2500 iteration (implemented using Stan and the R package *rstan*, Stan Development Team 2016). Initial starting values of parameters for the chains were set to a positive value determined by the priors for these parameters and some added randomness between chains, which was drawn from a normal distribution around 0 with a variation equal to 1/10$^{\text{th}}$ the prior standard deviation for these parameters. Exceptions were the taxon-specific parameters α_j, for which the starting values were determined by the relative frequency of occurrence of each taxon in the monitoring data, and some randomness between chains which was drawn from a normal distribution around 0 with a standard

deviation equal to 1/10th of the absolute value of the average relative frequency of occurrence of all taxa. All Bayesian models converged well.

S1.7. Model evaluation: standardized deviances and D^2 metrics

The standardized deviances for each taxon, d_j, and across all taxa, d, are described by:

$$d_j(\boldsymbol{\theta}) = -\frac{2}{n_j} \sum_{i=1}^{I} \sum_{t_i=1}^{T_i} \log\left(P\left(y_{it_ij}|\boldsymbol{\theta}\right)\right) \quad \text{(Eq. S1.7a)}$$

$$d(\boldsymbol{\theta}) = -\frac{2}{n} \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{t_i=1}^{T_i} \log\left(P\left(y_{it_ij}|\boldsymbol{\theta}\right)\right) \quad \text{(Eq. S1.7b)}$$

Here, n_j 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 is described by

$$D_j^2(\boldsymbol{\theta}) = \frac{d_j^{\text{null model}}(\boldsymbol{\theta}) - d_j^{\text{proposed model}}(\boldsymbol{\theta})}{d_j^{\text{null model}}(\boldsymbol{\theta})} \quad \text{(Eq. S1.7c)}$$

The D^2 statistic describes the fraction of the deviance of the null model that is reduced by the model. The null model in this context contains only the taxon-specific parameter α_j (β_r and γ_i are equal to zero) and assumes no influence of environmental factors. It corresponds to assuming a probability equal to the overall observed frequency of occurrence that is independent of the site.

S1.8. Stan code

In the first code segment, we define a function "interpolate" for 1-dimensional linear interpolation. This function will later be used to interpolate temperature, flow velocity, and saprobity values.

```
functions {

  // function for one-dimensional interpolation of x_points and y_points at x
  // ------------------------------------------------------------------------
  // x_points must be non-decreasing, y_points of the same length as x_points,
  // if x is outside the range of x_points, the first resp. last point of y_points is returned

  real interpolate(real x, vector x_points, vector y_points) {
    int n;
    n = num_elements(x_points);
    if ( num_elements(y_points) != n ) reject("*** interpolate: length of x  and y_points not equal");
    if ( x_points[n] < x_points[1] ) reject("*** interpolate: x_points must be non-decreasing ***");
    if ( x <= x_points[1] ) return y_points[1];
    if ( x >= x_points[n] ) return y_points[n];
    for ( i in 2:n ) {
      if ( x_points[i] < x_points[i-1] ) reject("*** interpolate: x_points must be non-decreasing");
      if ( x >= x_points[i-1] && x <= x_points[i] ) {
        if ( x_points[i-1] == x_points[i] ) return(0.5*(y_points[i-1]+y_points[i]));
        else return( ( (x-x_points[i-1])*y_points[i] + (x_points[i]-x)*y_points[i-1] ) /
                                                          ( x_points[i]-x_points[i-1] ) );
      }
    }
    reject("*** interpolate: adjacent x_points not found ***");
    return 0;
  }
}
```

The data block is used to transfer data from the R environment to stan. In R, a named list of variables has to be prepared that has to correspond to the stan definition in the data block. We first define the dimensions used later in various arrays and as index bounds in loops. This is followed by a section that transfers the trait information that is not estimated, such as various trait class boundaries, and the parameters that characterize the priors of estimated parameters (names appended by "`_pripar`"). The final variables `x` and `y` are the environmental factors and the data. The environmental factors, `x`, have one more column to transfer the waste water fraction of discharge that is used as an additional pesticide source indicator to the environmental factor of insecticide pollution. Note that because stan does not support data that is not available, the occurrence data in matrix `y` is encoded by –1 for missing values, 0 for absence, and +1 for presence of a taxon in a given sample. For this reason, we use an integer matrix bound to the range from –1 to +1, and we will later on exclude the negative data from the analysis.

```
data {
  int                       n_samples;
  int                       n_sites;
  int                       n_regions;
  int                       n_taxa;
  int                       n_trait;
  int                       n_pred;
  real                      mu_U_IAR_pripar;
  int                       trait_T_n;
  vector[2*trait_T_n]       trait_T_x;
  matrix[n_taxa,trait_T_n]  mu_trait_T_pripar;
  int                       trait_v_n;
  vector[2*trait_v_n]       trait_v_x;
  matrix[n_taxa,trait_v_n]  mu_trait_v_pripar;
  int                       trait_sap_n;
  vector[2*trait_sap_n]       trait_sap_x;
  matrix[n_taxa,trait_sap_n]  mu_trait_sap_pripar;
  vector[n_taxa]            mu_trait_pest_pripar;
  real                      mu_trait_pest_Kinvmax_pripar;
  vector[n_taxa]            mu_trait_morph_pripar;
  int                       trait_subst_n;
  matrix[n_taxa,trait_subst_n] mu_trait_subst_pripar;
  real                      mu_alpha_comm_pripar;
  real                      sigma_alpha_comm_pripar;
  vector[n_trait]           mu_beta_trait_pripar;
```

```
  vector[n_trait]           sigma_beta_trait_pripar;
  real                      fact_sd;
  real                      sigma_habsuit_pripar;
  int                       regionIND[n_samples];
  real                      x[n_samples,n_pred+1];
  int<lower=-1,upper=1>     y[n_samples,n_taxa];
}
```

Next, we have to specify the model parameters. `alpha` and `beta` correspond to the parameters α and β in Eq. 2. The trait parameters (`trait_`) correspond to the information from the trait databases except for the parameters `trait_pest_Kinvmax`, and `U_IAR,` which correspond to K_{invmax} defined in Eq. S1.4.2c and U_{IAR} described in Table S1.2. These 2 parameters belong to the parameter vector $\boldsymbol{u}_r$ in Eq. 1 in the main text.

```
parameters {
  matrix[n_taxa, n_regions]                   alpha_taxa;
  real<lower=0>                               beta_T;
  real<lower=0>                               beta_v;
  real<lower=0>                               beta_sap;
  real<lower=0>                               beta_pest;
  real<lower=0>                               beta_morph;
  real<lower=0>                               beta_subst;
  matrix<lower=0,upper=1>[n_taxa,trait_T_n]   trait_T_par;
  matrix<lower=0,upper=1>[n_taxa,trait_v_n]   trait_v_par;
  matrix<lower=0,upper=1>[n_taxa,trait_sap_n] trait_sap_par;
  vector<lower=0,upper=1>[n_taxa]             trait_pest;
  real<lower=1e-6>                            trait_pest_Kinvmax;
  vector<lower=0,upper=1>[n_taxa]             trait_morph;
  matrix<lower=0,upper=1>[n_taxa,trait_subst_n] trait_subst_par;
  real                                        U_IAR;
}
```

Finally, after the declaration of some local variables, the following block of code contains the definitions of all priors (section "root nodes"), of the internal nodes for the hierarchical parameters α (section "intermediate nodes"), and of a Bernoulli distribution for the observations, `y`, the parameter of which is calculated as an inverse logistic transformation according to Eq. 2 of the variable `z` calculated according to Eq. 3. Note that we exclude negative values of `y`, as –1 was the code for missing observations, as mentioned above.

```
model {

  // local variables:

  matrix[n_taxa,2*trait_T_n] trait_T_y;
  matrix[n_taxa,2*trait_v_n] trait_v_y;
  matrix[n_taxa,2*trait_sap_n] trait_sap_y;
  matrix[n_samples,n_taxa]     habsuit_T;
  matrix[n_samples,n_taxa]     habsuit_v;
  matrix[n_samples,n_taxa]     habsuit_sap;
  matrix[n_samples,n_taxa]     habsuit_pest;
  matrix[n_samples,n_taxa]     habsuit_morph;
  matrix[n_samples,n_taxa]     habsuit_subst;
  real                         s_pripar;
  real                         s;
  matrix[n_samples,n_taxa]     z;
  matrix[n_samples,n_taxa]     p;

  // root nodes:

  beta_T ~ normal(mu_beta_trait_pripar[1],sigma_beta_trait_pripar[1]);
  beta_v ~ normal(mu_beta_trait_pripar[2],sigma_beta_trait_pripar[2]);
  beta_sap ~ normal(mu_beta_trait_pripar[3],sigma_beta_trait_pripar[3]);
  beta_pest ~ normal(mu_beta_trait_pripar[4],sigma_beta_trait_pripar[4]);
  beta_morph ~ normal(mu_beta_trait_pripar[5],sigma_beta_trait_pripar[5]);
  beta_subst ~ normal(mu_beta_trait_pripar[6],sigma_beta_trait_pripar[6]);
  s_pripar = sqrt(log(1+fact_sd^2));
  U_IAR   ~ lognormal(log(mu_U_IAR_pripar)-0.5*s_pripar^2,s_pripar);
  for ( j in 1:n_taxa ) {
    for ( k in 1:trait_T_n ) {
      if ( mu_trait_T_pripar[j,k] >= 0 ) {
        trait_T_par[j,k] ~ normal(mu_trait_T_pripar[j,k],sigma_habsuit_pripar);
```

```
    }
  }
}
for ( j in 1:n_taxa ) {
  for ( k in 1:trait_v_n ) {
    if ( mu_trait_v_pripar[j,k] >= 0 ) {
      trait_v_par[j,k] ~ normal(mu_trait_v_pripar[j,k],sigma_habsuit_pripar);
    }
  }
}
for ( j in 1:n_taxa ) {
  for ( k in 1:trait_sap_n ) {
    if ( mu_trait_sap_pripar[j,k] >= 0 ) {
      trait_sap_par[j,k] ~ normal(mu_trait_sap_pripar[j,k],sigma_habsuit_pripar);
    }
  }
}
for ( j in 1:n_taxa ) {
  if ( mu_trait_pest_pripar[j] >= 0 ) {
    trait_pest[j] ~ normal(mu_trait_pest_pripar[j],sigma_habsuit_pripar);
  }
}
for ( j in 1:n_taxa ) {
  if ( mu_trait_morph_pripar[j] >= 0 ) {
    trait_morph[j] ~ normal(mu_trait_morph_pripar[j],sigma_habsuit_pripar);
  }
}
for ( j in 1:n_taxa ) {
  for ( k in 1:trait_subst_n ) {
    if ( mu_trait_subst_pripar[j,k] >= 0 ) {
      trait_subst_par[j,k] ~ normal(mu_trait_subst_pripar[j,k],sigma_habsuit_pripar);
    }
  }
}
s_pripar = sqrt(log(1+fact_sd^2));
trait_pest_Kinvmax ~ lognormal(log(mu_trait_pest_Kinvmax_pripar)-0.5*s_pripar^2,s_pripar);

// intermediate nodes:

for ( j in 1:n_taxa ) {
   for ( b in 1:n_regions ) {
       alpha_taxa[j,b] ~ normal(mu_alpha_comm_pripar,sigma_alpha_comm_pripar);
   }
}

// end nodes:

for ( j in 1:n_taxa ) {
  for ( k in 1:trait_T_n ) {
    trait_T_y[j,2*k-1] = trait_T_par[j,k];
    trait_T_y[j,2*k]   = trait_T_par[j,k];
  }
  for ( k in 1:trait_v_n ) {
    trait_v_y[j,2*k-1] = trait_v_par[j,k];
    trait_v_y[j,2*k]   = trait_v_par[j,k];
  }
  for ( k in 1:trait_sap_n ) {
    trait_sap_y[j,2*k-1] = trait_sap_par[j,k];
    trait_sap_y[j,2*k]   = trait_sap_par[j,k];
  }
  for ( i in 1:n_samples ) {
    habsuit_T[i,j] = interpolate(x[i,1],trait_T_x,trait_T_y[j,]');
    habsuit_v[i,j] = interpolate(x[i,2],trait_v_x,trait_v_y[j,]');
    habsuit_sap[i,j] = interpolate(x[i,3],trait_sap_x,trait_sap_y[j,]');
    habsuit_pest[i,j] = 1 / ( 1 + ((x[i,4] + U_IAR*x[i,13])*trait_pest[j]*trait_pest_Kinvmax)^2 );
    habsuit_morph[i,j] = x[i,5] * trait_morph[j] ;
    habsuit_subst[i,j] = 0;
    for ( r in 1:trait_subst_n ) {
      habsuit_subst[i,j] = habsuit_subst[i,j] + x[i,4+r]*trait_subst_par[j,r];
    }
  }
}
for ( j in 1:n_taxa ) {
  for ( i in 1:n_samples ) {
    z[i,j] = alpha_taxa[j,regionIND[i]];
```

```
      z[i,j] = z[i,j] + habsuit_T[i,j] * beta_T;
      z[i,j] = z[i,j] + habsuit_v[i,j] * beta_v;
      z[i,j] = z[i,j] + habsuit_sap[i,j] * beta_sap;
      z[i,j] = z[i,j] + habsuit_pest[i,j] * beta_pest;
      z[i,j] = z[i,j] + habsuit_morph[i,j] * beta_morph;
      z[i,j] = z[i,j] + habsuit_subst[i,j] * beta_subst;
      p[i,j] = 1/(1+exp(-z[i,j]));
      if( y[i,j] >= 0 ) {
        y[i,j] ~ bernoulli(p[i,j]);
      }
    }
  }
}
```

S1.9. References in Appendix S1

Bernatowicz, W., A. Weiss, and J. Matschullat. 2009. Linking biological and physicochemical water quality. Environmental Monitoring and Assessment 159:311–330.

Brooks, S., A. Gelman, G. Jones, and X. L. Meng. 2011. Handbook of Markov Chain Monte Carlo, CRC Press, Boca Raton, Florida.

Chow, V. T. 1959. Open-channel hydraulics. New York, McGraw-Hill Book Co., New York, New York.

Cowan, W. L. 1959. Estimating hydraulic roughness coefficients. Agricultural Engineering 37:473–475.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth. 1987. Hybrid Monte Carlo, Physics Letters B 195:216–222.

Holling, C. S. 1959. The components of predation as revealed by a study of small mammal predation of the European pine sawfly. The Canadian Entomologist 91:293–320.

Langhans, S. D., V. Hermoso, S. Linke, S. E. Bunn, and H. P. Possingham. 2013. Cost-effective river rehabilitation planning: Optimizing for morphological benefits at large spatial scales. Journal of Environmental Management 132c:296–303.

Langhans, S. D., J. Lienert, N. Schuwirth, P. Reichert. 2013. How to make river assessments comparable: A demonstration for hydromorphology. Ecological Indicators 32:264–275.

Liechti, P. 2010. Methoden zur Untersuchung und Beurteilung der Fliessgewässer: Chemisch-physikalische Erhebungen, Nährstoffe. Umwelt-Vollzug. Bundesamt für Umwelt, Bern, Switzerland. (Available from: https://www.modul-stufen-konzept.ch/download/ChemieD_Juni2010.pdf)

Moog, O. 2002. Fauna Aquatica Austriaca. Lieferung 2002. Wasserwirtschaftskataster, Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Vienna, Austria.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

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

Stucki, P. 2010. Methoden zur Untersuchung und Beurteilung der Fliessgewässer. Macrozoobenthos – Stufe F (flächendeckend), Umwelt-Vollzug Nr. 1026. Bundesamt für Umwelt, Bern, Switzerland. (Available from: https://www.modul-stufen-konzept.ch/download/MZB_Stufe_F-D_20111215.pdf)

Vermeiren, P., P. Reichert, N. Schuwirth. 2020. Integrating uncertain prior knowledge regarding ecological preferences into multi-species distribution models: Effects of model complexity on predictive performance. Ecological Modelling 420:108956

Wittmer, I. K., H. P. Bader, R. Scheidegger, H. Singer, A. Lück, I. Hanke, C. Carlsson, C. Stamm. 2010. Significance of urban and agricultural land use for biocide and pesticide dynamics in surface waters. Water Research 44:2850–2862

Zelinka, M., P. Marvan. 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. Archiv fur Hydrobiolgie 57:389–407.

S1.10. References to datasets in Appendix S1

BFS (2008a). Arealstatistik Schweiz: Zustand und Entwicklung der Landschaft Schweiz. Ausgabe 2008. Technical Report 897-0800, Bundesamt für Statistik BFS, CH-2010 Neuchâtel, Switzerland.

BFS (2008b). Landwirtschaftliche Betriebszählung / Census of farming. Technical report, Bundesamt für Statistik BFS, CH-2010 Neuchâtel, Switzerland.

FOEN (2013) MQ_GWN_CH: Mean runoff and flow regime types for the river network of Switzerland. Dataset, Federal Office for the Environment. (Available from: https://www.bafu.admin.ch/bafu/en/home/topics/water/state/maps/mean-monthly-and-annual-runoff/mean-runoff-and-flow-regime-types-for-the-river-network-of-switz.html)

Swisstopo. 2016. swissTLM3d. Technical Report Art. 30 Geo IV, swisstopo, Bern, Switzerland. (Available from: https://shop.swisstopo.admin.ch/en/products/landscape/tlm3D)

Appendix S2. HS-MSDM model performance and summary statistics

S2.1. Summary statistics for all taxa in HS-MSDM MS1

Table S2.1. Summary statistics for all taxa in the HS-MSDM MS1, including their relative occurrence and the D^2 and d statistics for model fit to the whole data and for testing against the 3 cross-validation runs.

➔ See separate file: App 2.1 Summary statistics per taxon.xlsx

S2.2. Overview of individual taxa predicted probabilities of occurrence in HS-MSDM MS1 and MF1.

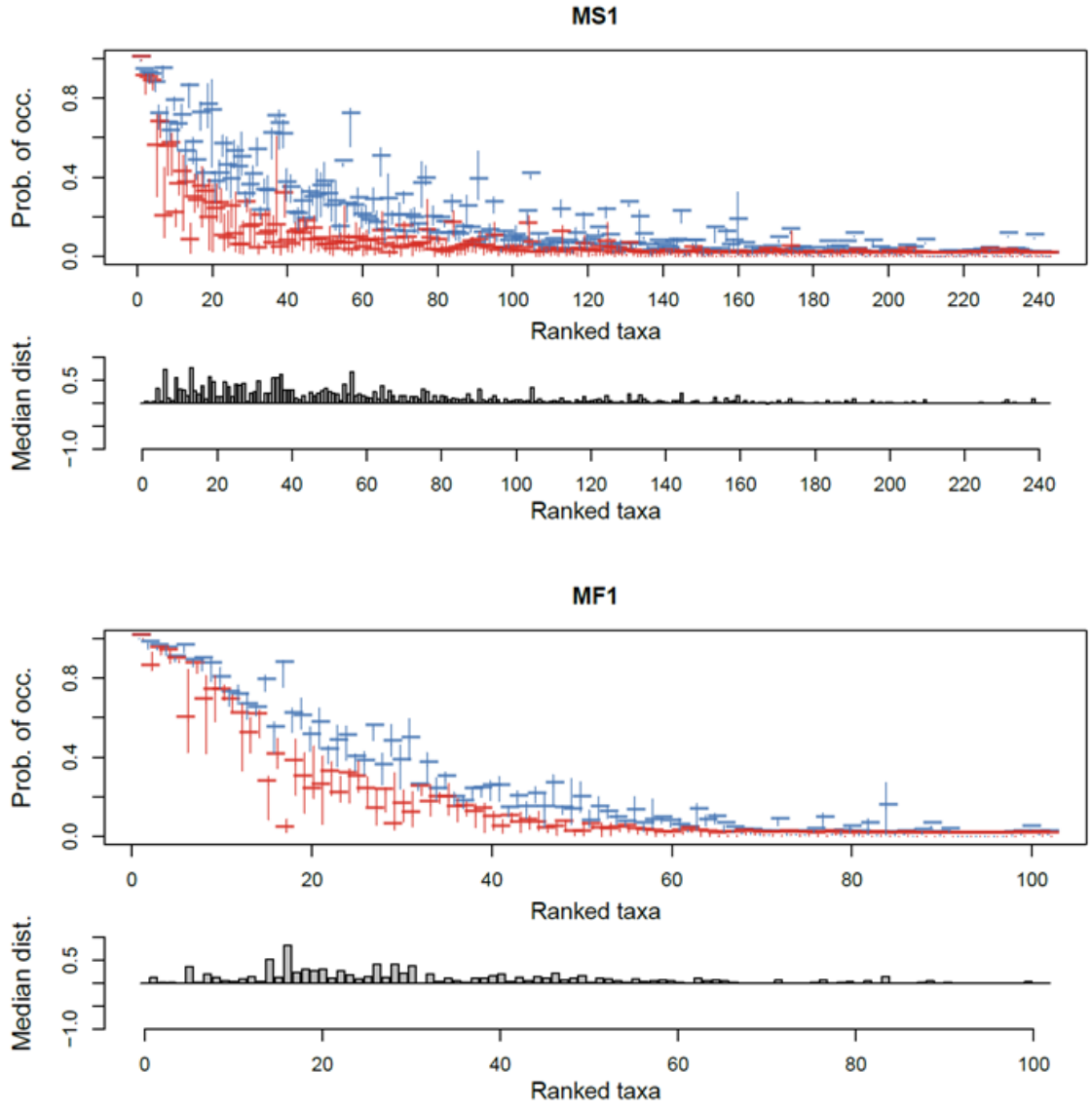

Fig. S2.2. Median (with 33–66 percentile) fitted probabilities of occurrence/taxon at sampling dates where the taxon is observed to be absent (in red) or present (in blue) in the top panel. The bottom panel shows the distance between the median fitted probabilities of occurrence/taxon between sampling dates where the taxon is observed as absent or as present. Taxa are ranked based on their relative frequency of occurrence, with the number of each taxon corresponding to their ranked order (see Appendix S2.1 for a full overview of individual taxa with their relative frequency of occurrence).

Appendix S3. Review of ecological preferences

S3.1. Prior-to-posterior shifts per ecological preference class

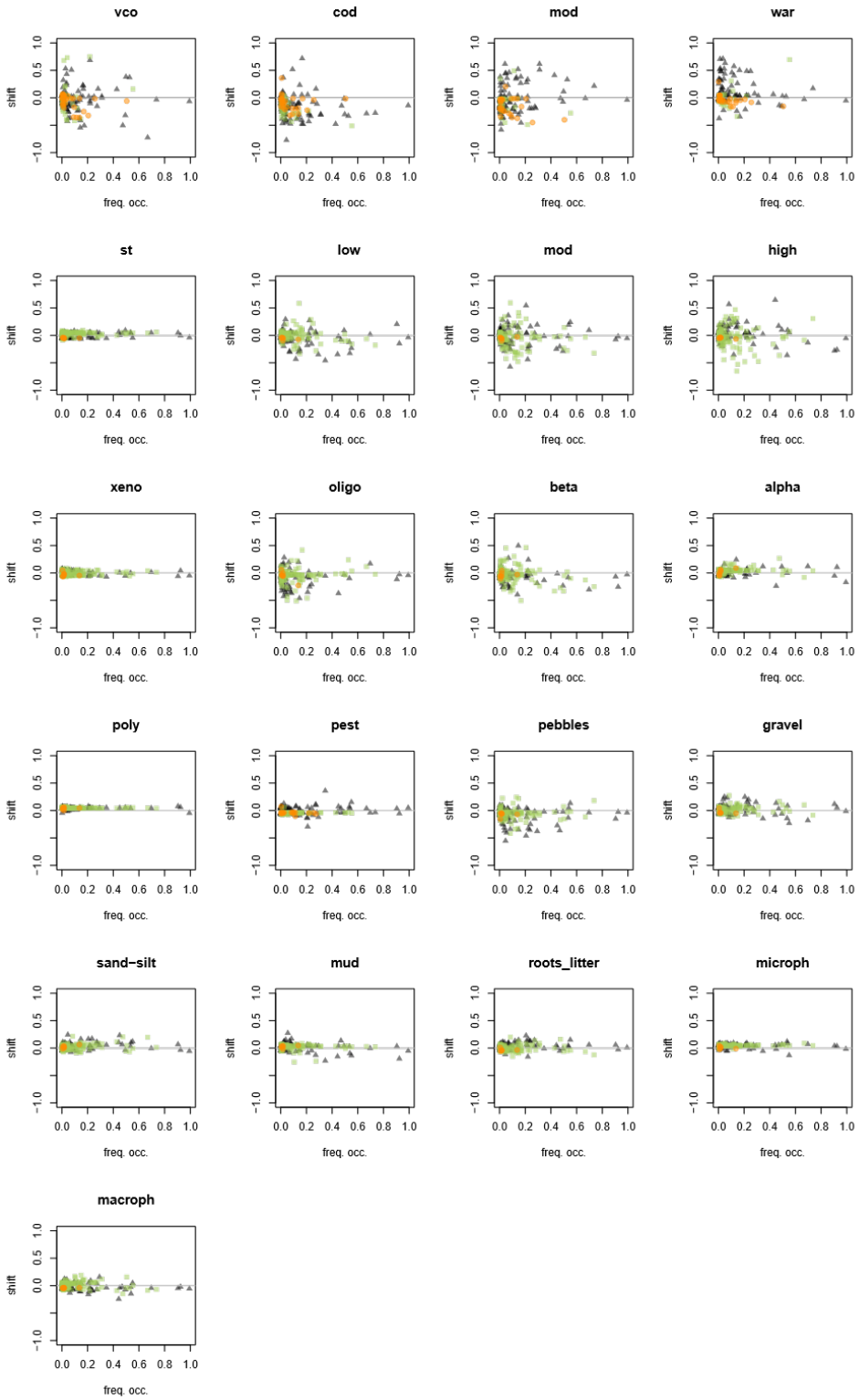

Fig. S3.1a Magnitude of prior-to-posterior shifts of ecological preferences for individual preference classes for each taxon (y-axis) vs the relative frequency of occurrence of the taxa (x-axis). Taxa for which prior information on ecological preferences was available are indicated as grey triangles, taxa that derived information from related taxa are indicated as green squares (derived from genus level) or orange circles (derived from family level). Results are for the MS1 model.

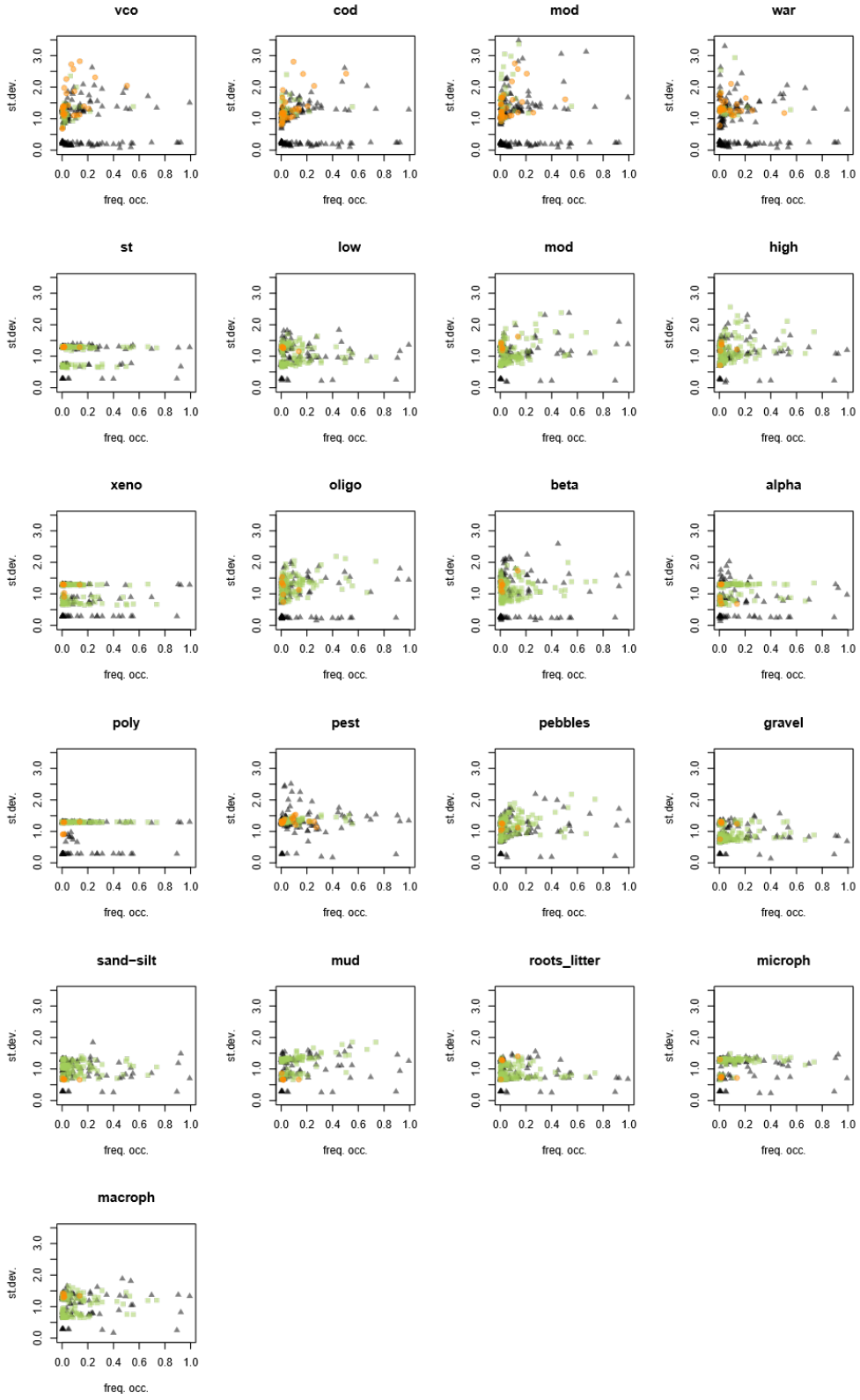

Fig S3.1b Plots of the standard deviation of the shifts in maximum prior to maximum marginal posterior parameter values for individual taxa for each trait class (on the y-axis) vs the relative frequency of occurrence of the taxa (x-axis). Species level taxa are indicated by grey triangles, genera by green squares, and families by orange circles.

S3.2. Table with suggested taxa to be revised

Table S3.2. Taxa for which the model suggests revising the ecological preferences; we only show taxa with a relative frequency of occurrence ≥0.15, a reasonable explanatory power of the model ($D^2 \geq 0.2$), and a prior-to-posterior shift of at least 0.2. The larger the shift, the more our posterior parameter estimates differ from the prior knowledge from the databases. Hence, the bigger the shift, the more our model suggests that the ecological preference is different (either higher if the number is positive or lower if the number is negative). We suggest that for these taxa, ecological experts review the plots showing for each ecological preference class the prior and posterior parameter distributions (Appendix S3.3, Fig. S3.3c) and assess whether a revision of the ecological preference scores are warranted (potentially using our maximum posterior parameter estimate as the new preference score).

	Temperature preference				Current preference				Saproby					Insecticides	Substrate preference						
	vco	cod	mod	war	st	slow	mod	high	xeno	oligo	beta	alpha	poly	pest	pebbles	gravel	sand.silt	mud	roots_litter	microph	macroph
Amphinemura		-0.32		0.49			-0.25	0.25													
Baetis alpinus		-0.28	0.22				-0.32	0.31			-0.25										
Baetis muticus	-0.5							-0.37			-0.27	0.27					0.21				
Baetis rhodani	-0.72	-0.29																			
Blephariceridae																					
Brachyptera risi				0.48		0.22		-0.25													
Cryptothrix nebulicola			-0.45																		
Dictyogenus		-0.29	0.33				-0.44														
Drusus discolor			-0.4																		
Ecdyonurus helveticus		-0.3	0.51																		
Ecdyonurus venosus	-0.51	-0.48					-0.22														
Elmidae								-0.36													
Epeorus alpicola	0.23	-0.41	0.27																		
Epeorus assimilis			-0.26	0.5			0.23	-0.31													
Gammaridae								-0.22		-0.37		-0.24			-0.27		0.23				
Habroleptoides confusa				0.25		-0.28												-0.24			
Leuctra braueri		-0.48	0.41								-0.33										
Leuctra major		0.72	-0.46	-0.34				-0.25		0.42	-0.5										
Leuctra nigra								-0.37													
Nemoura minima		-0.31	0.62					-0.47			-0.24										
Nemoura mortoni	0.38	-0.3					-0.23				-0.33										
Protonemura brevistyla	0.69	-0.43					-0.28														
Protonemura lateralis	0.37	-0.33	0.33				-0.28			0.24											
Rhithrogena loyolaea		-0.33				0.29	-0.3														
Rhithrogena puthzi	0.22	-0.47	0.38				-0.35														
Rhyacophila hirticornis	0.75	-0.24	-0.49																		
Rhyacophila tristis		-0.51	-0.28	0.7																	
Sphaeriidae										-0.21					-0.25						
Tinodes			0.28	-0.21						-0.35				-0.29							

S3.3. Plots for individual taxa for which a review is suggested (external files)

Fig. S3.3a Presence (blue points) and absence (red points) of taxa throughout Switzerland, with the predicted probability of occurrence/absence of the HS-MSDM MS1 indicated by the point size. Taxa plotted are those for which the model suggests a revision of ecological preferences (see Table S3.2)

➔ See extra file: Appendix Fig S33a Individual taxa maps HS-MSDM MS1.pdf

Fig. S3.3b Presence (blue points) and absence (red points) of taxa with the predicted probability of occurrence/absence of the HS-MSDM MS1 on the y-axis and the environmental gradients considered on the x-axis. Trait profiles extracted from the trait databases, normalized, and turned into a habitat suitability function are plotted in black. Taxa plotted are those for which the model suggests a revision of ecological preferences (see Table S3.2)

➔ See extra file: Appendix Fig S33b presence-absence across environmental gradients HS-MSDM MS1.pdf

Figure S3.3c Comparison of prior and posterior distributions for ecological preferences of taxa. The red line indicates the normalized ecological preference, which was derived from a database. The grey area shows the prior distribution that reflects the prior knowledge about the ecological preference from the database and its uncertainty. The black line illustrates the posterior distribution obtained from Bayesian inference, updating the prior distribution by confrontation with independent data through model calibration with HSMSDM MS1. Taxa plotted are those for which the model suggests a revision of ecological preferences (see Table S3.2)

➔ See extra file: Appendix Fig S33c Taxa ecological preferences HS-MSDM MS1.pdf