Ecological Modelling 475 (2023) 110170

Contents lists available at ScienceDirect

Ecological Modelling

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

SDM profiling: A tool for assessing the information-content of sampled and unsampled locations for species distribution models

Charles J. Marsh [a,b,*], Yoni Gavish [a], Mathias Kuemmerlen [c], Stefan Stoll [d,e], Peter Haase [e,f], William E. Kunin [a]

[a] School of Biology, University of Leeds, Leeds, LS2 9JT, United Kingdom
[b] Department of Plant Sciences, University of Oxford, Oxford, United Kingdom
[c] Charlotte-von-Stein-Straße 12, 53177 Bonn, Germany
[d] Environmental Campus Birkenfeld, University of Applied Sciences Trier, Birkenfeld, Germany
[e] Faculty of Biology, University of Duisburg-Essen, Essen, Germany
[f] Department of River Ecology and Conservation, Senckenberg Research Institute and Natural History Museum Frankfurt, Gelnhausen, Germany

A R T I C L E I N F O

Keywords:
Active learning
Conservation
Ecological niche models
Model evaluation
Monitoring
Uncertainty

A B S T R A C T

Species distribution models (SDMs) are key tools in biodiversity and conservation, but assessing their reliability in unsampled locations is difficult, especially where there are sampling biases. We present a spatially-explicit sensitivity analysis for SDMs – SDM profiling – which assesses the leverage that unsampled locations have on the overall model by exploring the interaction between the effect on the variable response curves and the prevalence of the affected environmental conditions. The method adds a 'pseudo-presence' and 'pseudo-absence' to unsampled locations, re-running the SDM for each, and measuring the difference between the probability surfaces of the original and new SDMs. When the standardised difference values are plotted against each other (a 'profile plot'), each point's location can be summarized by four leverage measures, calculated as the distances to each corner. We explore several applications: visualization of model certainty; identification of optimal new sampling locations and redundant existing locations; and flagging potentially erroneous occurrence records.

1. Introduction

Knowledge of species' distributions is key to successful conservation measures, but in nearly all cases species are known from incomplete, and often spatially-biased, observations. Consequently, species distribution models (SDMs) have become key tools in biodiversity monitoring and conservation planning (Guisan et al., 2013). They allow for a set of presence only, presence/pseudo-absence or presence/absence data (hereafter, occurrence data) to be used to infer a species' distribution across the remainder of the unsampled region (Elith and Leathwick, 2009). Such outputs may then be used for a wide range of conservation applications, such as tracking changes in the distribution of target species to identify increasing or decreasing trends (e.g. Brotons et al., 2007), projecting potential future range-shifts (e.g. Elith et al., 2010), or to identify critical conservation areas essential for species persistence and reintroductions, or the protection of biodiversity (e.g. Kremen et al., 2008; Riaz et al., 2020).

For SDMs to be used effectively, it is therefore essential that such model outputs are accurate representations of the true distributions. The accuracies of SDMs are dependant not only upon the sampling effort (the quantity of occurrence data) used to generate the models (Aizpurua et al., 2015; Valavi et al., 2021), but also the spatial configuration of those sampling points (Kramer-Schadt et al., 2013; Syfert et al., 2013), particularly for presence-pseudoabsence models (Barbet-Massin et al., 2012; Phillips et al., 2009). As there is likely to be considerable sample selection bias in occurrence data, any SDM therefore risks conflating modelling species distribution with modelling this sampling bias (Beck et al., 2014; Phillips et al., 2009; Ploton et al., 2020; Radosavljevic and Anderson, 2014).

Furthermore, SDMs are generally evaluated through metrics summarising their accuracy against a set of validation data set-aside from the modelling procedure (Fielding and Bell, 1997), ensuring maximum fit to those locations with many data. This can be especially problematic if the validation data have the same spatial and environmental biases as the modelling data (Bahn and McGill, 2013; Leroy et al., 2018), such that a high level of accuracy to a small area in geographical or

* Corresponding author.
E-mail addresses: charlie.marsh@mailbox.org (C.J. Marsh), W.E.Kunin@leeds.ac.uk (W.E. Kunin).

https://doi.org/10.1016/j.ecolmodel.2022.110170
Received 13 June 2022; Received in revised form 27 September 2022; Accepted 7 October 2022
Available online 3 November 2022

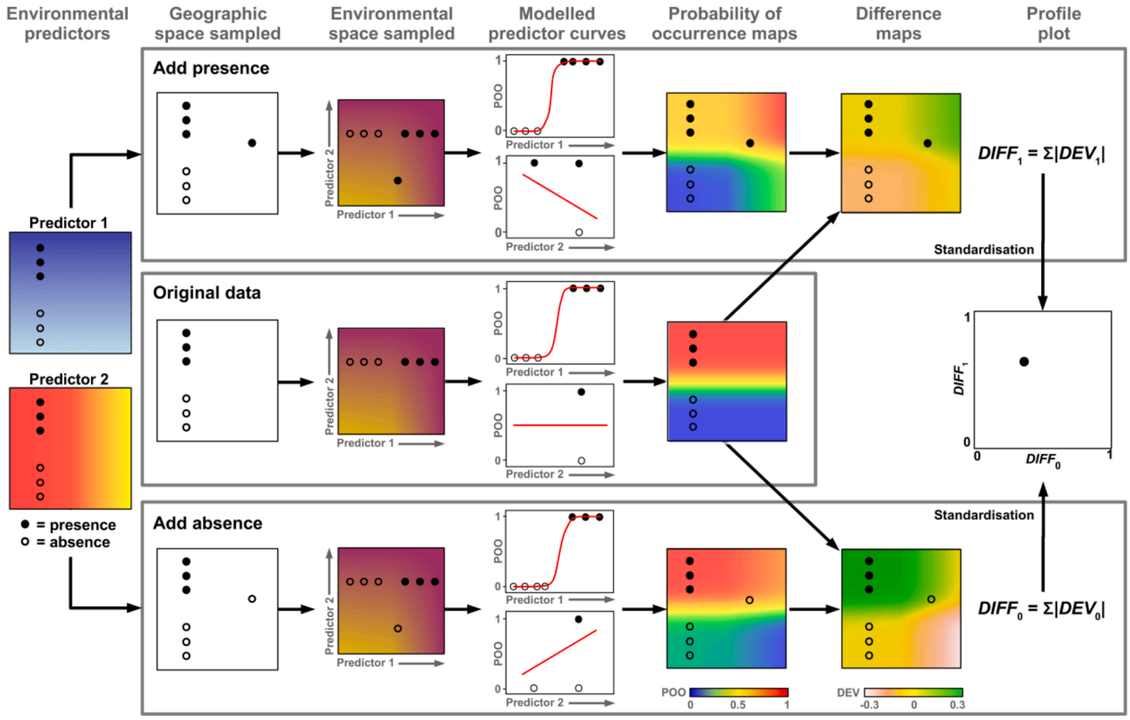

Fig. 1. A flow of the SDM profiling procedure. We start with six sampled locations which we model against two environmental predictors (column 1). We then add a new 'sampling' point (column 2) where we place a virtual presence (top row) or a virtual absence (bottom row). We have therefore 'sampled' a new area of environmental space for modelling (column 3), which in turn leads to different response curves for environmental predictor 2 in the model (column 4). The models are used to predict probability of occurrence across the area (column 5) and the differences for all cells calculated between the maps created with and without the new 'sampling' point. The summed difference when the 'sampling' point was a presence (*DIFF*$_1$) and an absence (*DIFF*$_0$) are then plotted against each other in a profile plot, and the procedure repeated for all other unsampled cells.

environmental space is assumed to apply to all areas beyond those bounds (Charney et al., 2021). Validation data can be carefully structured or filtered to minimize the influence of spatial biases as far as possible (e.g. Boria et al., 2014; Hallman and Robinson, 2020; Kramer-Schadt et al., 2013; Roberts et al., 2017), but any selection is still limited by where available samples are located. Although there is some movement towards producing maps of model uncertainty (Beale and Lennon, 2012; Rocchini et al., 2011; Swanson et al., 2013) estimating accuracy in the model outputs in a spatially-explicit manner is still challenging.

Here we develop such a method for evaluating SDMs, which we call SDM profiling, that highlights key locations with the potential to affect model predictions should further data be collected. It does this by calculating the leverage that unsampled areas would have on the overall model predictions (rather than simply the effect on the response curves themselves) if they were to be sampled. We first describe the SDM profiling procedure in detail. We then show that the values generated for each unsampled cell provide a meaningful estimate of its likely effect on our model predictions using simulated species distributions. Finally, using simulated species and data for a freshwater gastropod from a long-term monitoring scheme, we highlight several potential applications for SDM profiling, including as a tool for visualising areas of high and low model certainty, selecting sites for collecting further data, optimising monitoring schemes and flagging potentially erroneous occurrence records.

2. Materials and methods

2.1. The SDM profiling procedure

The basic assumption of SDM profiling is that each cell (sampled or unsampled) can be assigned to one of only two potential states: presence or absence. Thus, we can explore the effect of the cell on an SDM by comparing the predicted probability of occurrence map when the cell is included in the analysis to a probability of occurrence map produced when the cell is not included in the analysis. For example, we can 'virtually' sample an unsampled cell by including it in the SDM training data in each of its two possible states, once when the cell is assumed to be a presence and once when assumed to be an absence, to generate two new probability of occurrence maps (Fig. 1). These two measures of change can be standardized and plotted against each other in a 'profile plot'. The location of the cell in this plot relative to other cells allows the quantification of the information leverage of each cell for both possible states.

The method incorporates not only information on the change to the response curves of the environmental variables if a cell were to be sampled, but also the prevalence of those conditions within the landscape (Ewers et al., 2010). It therefore differs fundamentally from methods that only examine which unsampled locations would cause the largest changes to the response curves, such as traditional statistical leverage or influence measurements like the Cook's distance in a generalized-linear model. For example, a new sampling location might cause a very large change in predicted probability of occurrence in a particular portion of an environmental gradient. If, however, a very small portion of the landscape possess those environmental conditions then the effect on the overall accuracy of our predictions will be small.

Since SDM profiling always compares a new probability of occurrence map to an original one, it can be applied to any species for which we have some initial occurrence data that can be used to create a base SDM. As the procedure directly compares the probability of occurrence values generated by multiple SDMs it is therefore advisable to build models using true presence-absence data, where detectability issues are accounted for, rather than presence only models that may predict only relative likelihoods (Guillera-Arroita et al., 2015).

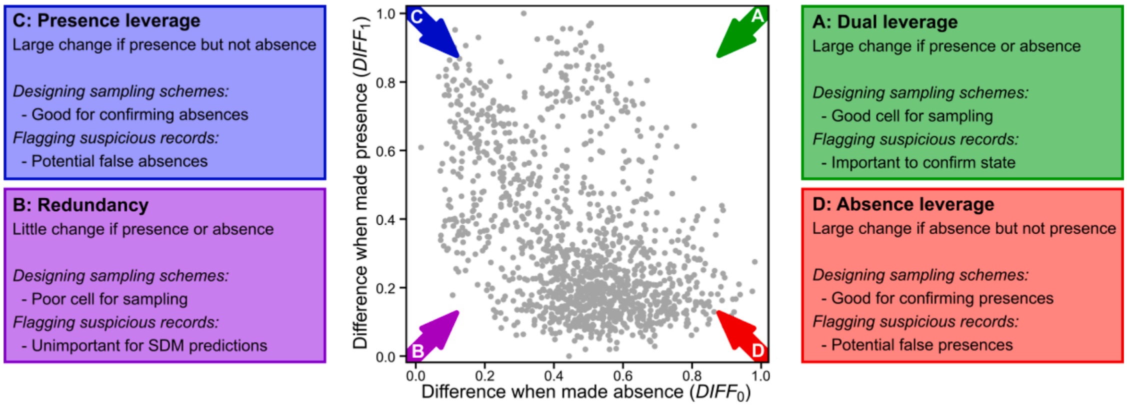

Fig. 2. A profile plot showing the overall standardised change in modelled probabilities of occurrence when each cell is made a presence ($DIFF_1$) and an absence ($DIFF_0$). We identify four measures of information leverage calculated as the distance to each plot corner: (A) dual-leverage (green), (B) redundancy (purple), (C) presence-leverage (blue) and (D) absence-leverage (red). The four measures may provide valuable information for different applications (see boxes).

To start the procedure we first use the initial occurrence data to fit an SDM and create the base probability of occurrence map to which all other generated maps will be compared. At this stage we have for each cell j (from a total of J cells in the map), the predicted probability of occurrence according to the base SDM (hereafter BASE_j). We then generate a list of all the K cells we wish to profile (a subset of J). For a given cell k (in K) we set its state to presence and re-run the SDM to produce a new probability of occurrence for each cell ($NEW_{1,k,j}$). We then repeat this procedure for the same cell k, while setting its state as absence to produce a new probability of occurrence for each cell j ($NEW_{0,k,j}$).

For each cell j we calculate the absolute deviation from the probability of occurrence in the base SDM with the new probability of occurrence when cell k is changed to a presence or absence:

$$DEV_{1,k,j} = \left|NEW_{1,k,j} - BASE_j\right| \tag{1}$$

$$DEV_{0,k,j} = \left|NEW_{0,k,j} - BASE_j\right| \tag{2}$$

Then, for each cell k, the $\mathrm{DEV}_{1,k,j}$ and $\mathrm{DEV}_{0,k,j}$ are summed over all j cells to represent the total change in probability of occurrence across all cells when cell k is made a presence or an absence respectively. Next, the maximum and minimum deviation across all cells are calculated:

$$\Delta_{max} = max_k\left(\sum_{j=1}^{J} DEV_{0,k,j}, \sum_{j=1}^{J} DEV_{1,k,j}\right) \tag{3}$$

$$\Delta_{min} = min_k\left(\sum_{j=1}^{J} DEV_{0,k,j}, \sum_{j=1}^{J} DEV_{1,k,j}\right) \tag{4}$$

Finally, for each focal cell k, the total deviation is standardized between 0 and 1 by the maximum and minimum deviations to calculate the total change in probability of occurrence when assumed presence or absence:

$$DIFF_{1,k} = \frac{\sum_{j=1}^{J} DEV_{1,k,j} - \Delta_{min}}{\Delta_{max} - \Delta_{min}} \tag{5}$$

$$DIFF_{0,k} = \frac{\sum_{j=1}^{J} DEV_{0,k,j} - \Delta_{min}}{\Delta_{max} - \Delta_{min}} \tag{6}$$

2.2. The 'profile plot' and the four measures of leverage

When repeated for all K cells we can then plot the standardised $DIFF_{1,k}$ and $DIFF_{0,k}$ values against each other in a 'profile plot' (Fig. 2). The position of cell k in this plot is therefore an indication of the total information leverage that cell would have on the model if it was sampled and the species was found to occur there or not.

We recognise four leverage measures that can be derived from the plot relating to the proximity to each of the four corners. In all cases we calculate proximity as $\sqrt{2}$ minus the Euclidean distance to the corner, such that the largest leverage values for a given corner indicate the closest points. These four measures are heuristically useful, even though they overspecify the data; once any three are known, the location of any point could be triangulated, making the fourth variable redundant. Indeed, the four measures can be categorised along two axes.

2.2.1. Leverage strength

The first axis runs from corner A to B in Fig. 2 and relates category to 'leverage strength'. The further along this axis towards corner A the stronger the overall leverage of that cell.

- A *Dual-leverage*— Cells in this corner indicate locations that have high leverage whether the species is present or absent. This may occur for example for cells in poorly sampled and/or prevalent environments where small changes in the modelled responses can result in large sum changes in probability across the landscape. We cannot therefore make a-priori judgements on the species' likely state in that cell, but know that the cell has important consequences for our model predictions. Cells with high dual-leverage values are therefore good candidates for future sampling.
- B *Redundancy*— For cells close to this corner, although leverage is similar whether made a presence or absence, the change in model predictions are small in both cases and so they are not likely to be providing information not already present in the model. This may occur if the environmental conditions are rare in the landscape or already well sampled. Cells with high redundancy values would therefore be inefficient selections for future sampling.

2.2.2. Leverage symmetry

The second axis runs for corner C to D and relates category to 'leverage symmetry'. Location along this axis indicates that leverage is stronger for one state (i.e. either presence or absence) than the other.

- C *Presence-leverage*— Cells near this corner indicate that there is a large change in the predicted probability of occurrence if the species were to occur in that cell, but conversely the predicted probability of occurrence remains largely unchanged when the cell is made an absence. It is therefore an unlikely site for the occurrence of the species if those environmental conditions have been well sampled. For example, if modelling a species that is confined only to woodland, adding an absence to a grassland cell would not greatly affect our predictions, but adding a presence to that cell may lead to

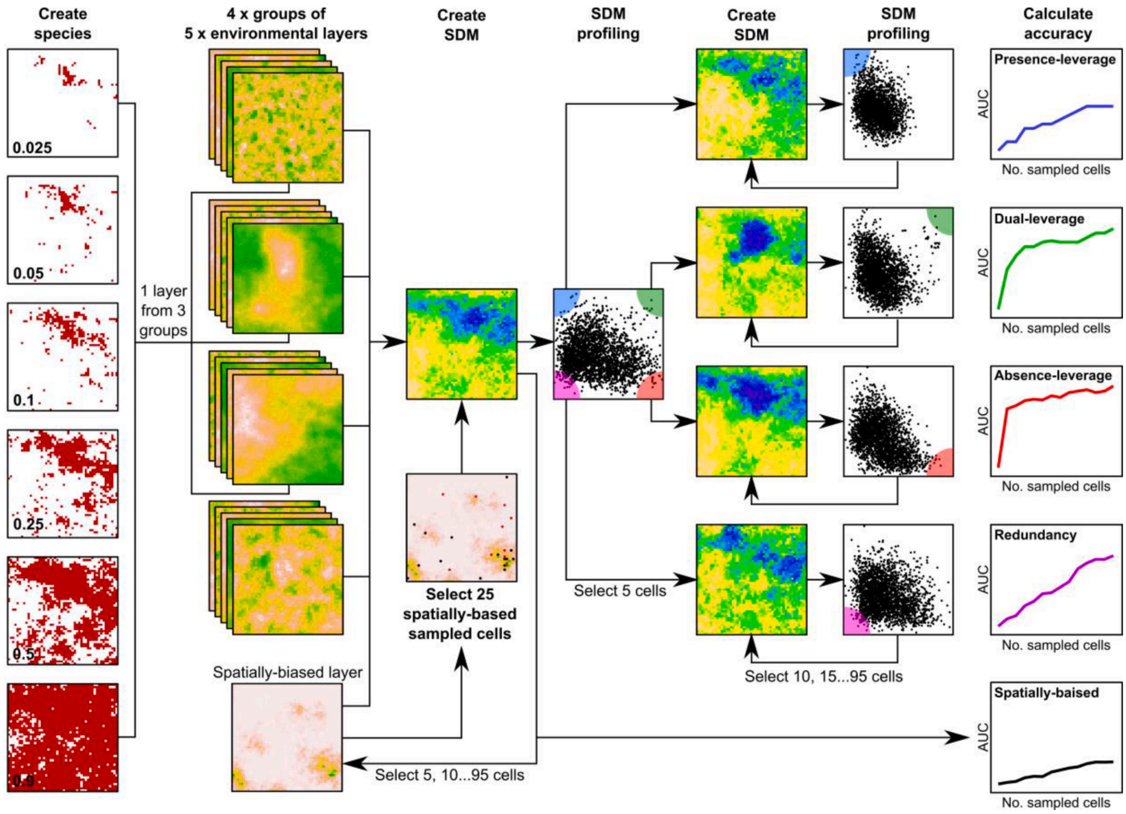

Fig. 3. Outline of the simulation procedure for a set of species with six prevalence values. This procedure was repeated 30 times each where species had aggregated or dispersed distributions.

increases in suitability across all grassland cells, especially if grassland has hitherto been poorly sampled. Therefore, sampling at such a site may further confirm what the model currently predicts as absences if those conditions are well-sampled, but greatly change the model if those conditions were poorly sampled and modelled preference is based on relatively little data. The incorrect assignment of a presence in these cells has the potential to significantly mislead predictions.

D *Absence-leverage*— Cells in the opposite corner indicate a small leverage when a presence is added, but a large change in the model with the cell is made an absence. Again, this may indicate that the cell is deemed likely to be a presence, for example a woodland cell in the previous example. Here, the incorrect assignment of an absence has the potential to significantly mislead predictions.

It is also worth noting that the degree of change a cell's sampling will have on the model is a function of two factors. First, leverage is a function of the spatial and environmental information that the cell is providing for the model, so that a cell located in environmental conditions dissimilar from those already sampled will likely have higher leverage values than a cell located within a heavily-sampled region of environmental space. Second, leverage is also dependant upon the prevalence of those environmental conditions within the landscape. For example, the altered model may double the probability of occurrence for a certain environmental condition but if that condition is scarce within the landscape the overall change in the probability of occurrence across all cells will be small. Conversely, only a slight increase in probability of occurrence for those conditions prevalent in the landscape may lead to a large overall change. This highlights the general importance in spatial ecology that the effect sizes themselves may not be informative until mapped and aggregated across the landscape (Ewers et al., 2010), as well as explaining why simply selecting cells based on maximising the sampled environmental space is also likely to be an inefficient solution.

2.3. Testing the relevance of the four leverage measures

We illustrate the relevance of a cell's position in the profile plot, and the four leverage measures, through simulations (Fig. 3). We created a series of virtual species and generated SDMs from initially a small number of randomly assigned cells. Further cells were subsequently added iteratively by selecting cells closest to the respective corners of the profile plot and the change in accuracy of the SDM predictions calculated. By utilising virtual species we were able to assess the true accuracy of each SDM as we increased sampling effort. Consistent differences in the rate of increase in model accuracy for species of different prevalences indicate if our leverage measures do indeed reflect different informational content for unsampled cells.

All simulations were carried out in R 3.4.3 (R Core Team, 2020). Environmental and species layers were generated using the gstat package (Gräler et al., 2016; Pebesma, 2004), and SDMs were generated using the randomForest package (Liaw and Wiener, 2002), although SDM profiling could be applied to any SDM algorithm that outputs probability of occurrence values. An R package which contains functions for carrying out SDM profiling and plotting the results can be installed from https://github.com/charliem2003/sdmProfiling. The package also contains functions for creating simulated species and sets of environmental variables that follows the described methodology.

2.3.1. Creation of virtual environmental variables and species distributions

Environmental variables— Environmental variables were generated in groups of five across landscapes of 50 × 50 cells. First a surface was created using an unconditional Gaussian simulation to create a random field from a spherical variogram using a consistent, but arbitrary, sill value of 1.5. Spatial autocorrelation, controlled using the range parameter, was drawn from an exponential distribution (the exponential of a value drawn randomly between 1 and 6). Larger values result in higher levels of spatial autocorrelation. Four further environmental variables were then created based upon the first by subsampling 50% of the cells of the first variable. The subsampled cells were then used as input for a new Gaussian simulation with new spatial autocorrelation values, randomly assigned from the same distribution. Finally, values for all five variables were standardised between 0 and 1. This resulted in a group of five closely-related variables, but each with different levels of spatial autocorrelation. We repeated this procedure a further three times to produce four groups of five environmental variables.

Only the first variable from three groups was used to generate the virtual species distributions themselves (see below), but the SDMs were trained with all twenty environmental variables as predictors. In real-world cases we generally don't know *a-priori* which environmental variables are important in determining the distribution of the species in question, and so usually there will be unimportant variables that are erroneously included. Our approach here means that there is ample scope for an SDM to incorrectly identify the important predictors and for this to lead to an inefficient accumulation of new sampling points.

Sampling bias— A further independent environmental variable with very high spatial autocorrelation (Range = 500) was created that approximates a typical spatial bias in sampling probability. We might expect similar real-world sampling probability surfaces (e.g. Beck et al., 2014) due to, for example, higher sampling effort around cities and universities (the "botanist effect"; Moerman and Estabrook, 2006). This layer was generated independently of any environmental or species layer.

Virtual species— To generate a virtual species for a given set of environmental variables, we multiplied the first variable of three of the environmental groups. We subsampled 50% of these cells and used them as input to generate a new surface layer in the same manner as for the environmental variables, using an unconditional Gaussian simulation to create a random field from a spherical variogram with a sill value of 1. The surface was then converted to a presence-absence distribution by selecting the number of highest value cells that equalled the desired proportion of occupied cells. The procedure was repeated for six species prevalence values for a given set of environmental variables, with proportions of 0.025, 0.05, 0.1, 0.25, 0.5 and 0.9 occupied cells so that for a given run all six species had similar underlying distributions but different prevalences.

In order to check that the species had appropriate distribution characteristics, we calculated the area-weighted mean class aggregation index value (He et al., 2000). As aggregation indices are dependant upon prevalence we only carried this out for prevalence of 0.1. Species were assigned as having an aggregated distribution if the aggregated index value fell between 80 and 90, and as having a dispersed distribution if the aggregated index value fell between 30 and 40. Although the thresholds are somewhat arbitrary, they produced subjectively very different distribution patterns (see examples in Fig. S1 in Supplementary Material). If the species could not be assigned to either then the species and environmental variables were discarded and the procedure restarted.

Initial sampling cells— For each species we selected 25 cells in order to build the initial SDMs. We assumed perfect detection so that models were built with presence-absence information. Cells were drawn from the sampling bias layer by generating a probability value for each cell from a uniform distribution between 0 and 1. Cells where the sampling bias value exceeded the probability value were assigned as 'sampled'. We then drew 25 of the 'sampled' cells at random, ensuring a minimum of three presences or absences.

Therefore, our knowledge of the species distribution starts from a spatially-biased position, with no correlation to environmental variables or the species distribution, approximating spatial biases that predominate in biological databases. We also present results in the supplementary material where the initial 25 cells were sampled randomly with no spatial bias.

2.3.2. Sampling procedure

For a given species within a set of environmental variables, we first generated an initial SDM using the 25 starting sampling cells as presence-absences. All SDMs were built using random forests with 2500 trees. Although many other algorithms are available for building SDMs, random forest models are very rapid to compute, which is important as our simulations required fitting some 123,651,480 models. However, the procedure should be robust to the modelling approach used if other algorithms are preferred. SDMs were built using all 20 environmental variables as well as the sampling bias layer as independent variables, including quadratic terms. As species distributions were built using only three environmental variables multiplied, we were therefore providing models that were greatly more complex than necessary and of high risk of overfitting if provided with spatially or environmentally-biased training data. Accuracy of the SDM prediction was then evaluated against the true species distribution through the true skills statistics (TSS) after applying a presence-absence threshold of 0.5. Accuracy was also assessed using the Kappa statistic and AUC and presented in the supplementary material.

Further sampling sites were added with increasing increments of sampling effort. Initially five further sampling cells were added, followed by a further 10 cells, 15 cells, to 95 cells in 5 cell increments for a total of 19 iterations. Therefore, sampling effort ranged from 25 to 975 cells.

For each iteration, a profile plot was generated for all unsampled cells. Additional sampling cells were selected based upon the four leverage measures generated from the profile plot: a) the highest dual-leverage values; b) the highest redundancy values; c) the highest presence-leverage values; d) and the highest absence-leverage values. As a null model, we also accumulated additional sampling cells based upon the same spatially-biased layer used for assigning the initial 25 sampled cells, without consideration of the species' potential distribution through SDMs, as is often typical in non species-specific sampling schemes or ad-hoc recording. In all cases, at each iteration a new SDM was generated and its accuracy assessed against the true distribution.

We replicated the procedure 30 times, generating 30 new sets of environmental variables for each species type (aggregated and dispersed) and 6 species prevalences for each environmental variable set. The success of the five approaches for accumulating sample cells was assessed as the increase in accuracy as sampling effort increased, averaged across the 30 replicates for each species. We also explored the relationship between the probability of occurrence of cells against the leverage values once the cells were profiled for a representative species. If correlations are high then it may indicate that there is little additional information gained from the profiling procedure. However, even where probability of occurrence and leverage are correlated, if there is variation in leverage values for a given probability of occurrence, then selecting cells based on probability of occurrence alone would potentially result in selecting cells with relatively small leverage values.

We explored several potential applications for SDM profiling, including visualization of model certainty in RGB colour space, identification of new sampling locations and redundant existing locations and the flagging potentially erroneous occurrence records. Applications were illustrated using either virtual species generated using the same procedure as above (the example species is the same shown in Fig. 6), where the initial sampling data comes from 25 sampling points arranged across a regular grid, or a real-world scenario of an existing, freshwater monitoring scheme, the Long-term Ecological Research (LTER; Mirtl et al., 2018) site Rhine-Main-Observatory (RMO; Kuemmerlen et al., 2016), focusing on data for a freshwater gastropod (*Ancylus fluviatilis* O. F. Müller, 1774). In this monitoring scheme, 21 existing sampling points were located along the main channel and a further 50 unsampled locations were identified as potential additions for future monitoring. RMO data is publicly available from https://rmo.senckenberg.de/search/home.php.

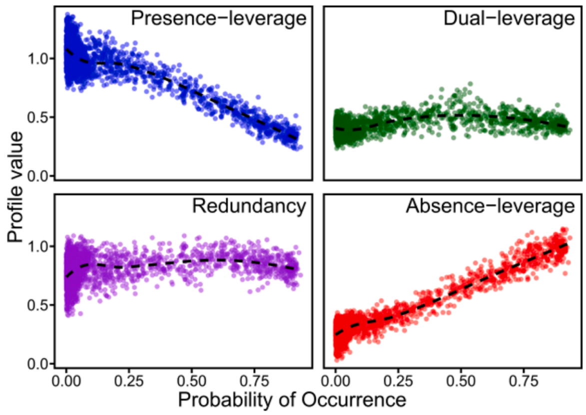

Fig. 4. The relationship between the probability of occurrence predicted for unsampled cells from a random forest SDM against the values of the four leverage measures of the profile plot, which measures the overall change of the SDM predictions if those cells were to be sampled: presence-leverage (blue), dual-leverage (green), redundancy (purple) and absence-leverage (red). Dashed lines are best-fit loess smoothers.

3. Results

When the profile values were plotted against the probability of occurrences of the SDM profiled (Fig. 4), leverage symmetry (presence- and absence-leverage values), but not leverage strength (redundancy and dual-leverage values), were highly correlated with probability of occurrence. Cells with high presence-leverage values have low probability of occurrence whereas cells with high absence-leverage have high probability of occurrence. There is some indication that those cells with the highest dual-leverage values have probability of occurrence values around 0.5, suggesting that the measure is focusing on cells of highest uncertainty, but importantly selecting sites on medium probability of occurrence alone gives little indication of their redundancy or dual-leverage value.

Selecting new sampling locations based upon their leverage values greatly affected our ability to accurately predict a species' distribution in consistent ways (Fig. 5), confirming the relevance of the profile plot. However, the identity of the most appropriate sampling method differed depending upon the prevalence of the species.

In general, selecting points using redundancy (Fig. 5; purple lines), those points that result in little change to the SDM output when added, led to slow increases in accuracy, whereas selecting points based on dual-leverage (Fig. 5; green lines), those points that cause large change to the SDM output when added, generally led to rapid increases in accuracy, especially when adding to small numbers of sampled cells or for species with medium prevalences. Accumulating samples based only on a spatial bias independent of the species (Fig. 5; orange lines) was inefficient across all scenarios. For prevalent species, accumulating sampling points based on presence-leverage (low values for $DIFF_0$; high values for $DIFF_1$; Fig. 5: blue lines) led to rapid increases in accuracy, whereas for rare species (i.e. the majority of species on earth), accumulating sampling points based on absence-leverage (high values for $DIFF_0$, low values for $DIFF_1$; Fig. 5: red lines) led to rapid increases in accuracy. For example, selecting points for a rare species based upon absence-leverage can almost double accuracy compared to other methods, whereas selecting points based upon presence-leverage would lead to only half as much increase in accuracy (Fig. 5). We can see in Fig. 6 (first column) that using absence-leverage concentrates points on the few true presences, and so even though the selected points covers only a small proportion of the environmental space (second column) the model can accurately predict the edges of the species range.

By contrast, selecting points for rare species through presence-leverage values identifies locations where there is high certainty of absence which are of little value to increasing model accuracy. Although a wide environmental space is sampled many cells need to be sampled before the environmental space dividing presences and absences has been adequately covered.

Selecting points based on redundancy (low values for $DIFF_0$; low values for $DIFF_1$) accumulates points in a spatially-clustered manner, whereas selecting points based upon dual-leverage (high values for $DIFF_0$; high values for $DIFF_1$) samples widely across the entire landscape, roughly in proportion to prevalence and so provides a good compromise between absence- and presence leverage whilst performing well across all species prevalences (Fig. 5).

4. Discussion

We can envisage three main uses for SDM profiling, ranging from model visualisation to optimising sampling schemes and flagging

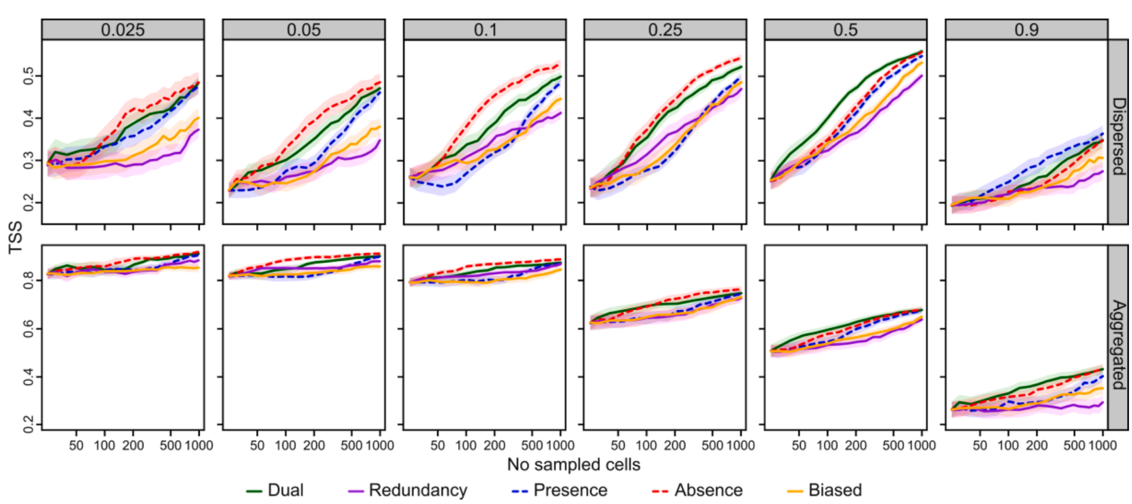

Fig. 5. The accuracy measured through the True Skills Statistic (TSS) for modelled distributions against sampling effort for six virtual species with prevalences of 0.025, 0.05, 0.1, 0.25, 0.5 and 0.9 and two distribution types: aggregated (bottom row) and dispersed (top row). Sites were sequentially added using four measures of selection based upon SDM profiling: dual-leverage (green), redundancy (purple), presence-leverage (blue dashed), absence-leverage (red dashed) and a spatial bias independent of the species distribution (orange). The initial 25 cells were drawn from a biased sampling surface independent of the species distribution. Results from iterations where initial sample sites were randomly drawn are available in the Supplementary Material. Solid lines are the mean from 30 replicates and polygons the 95% confidence interval.

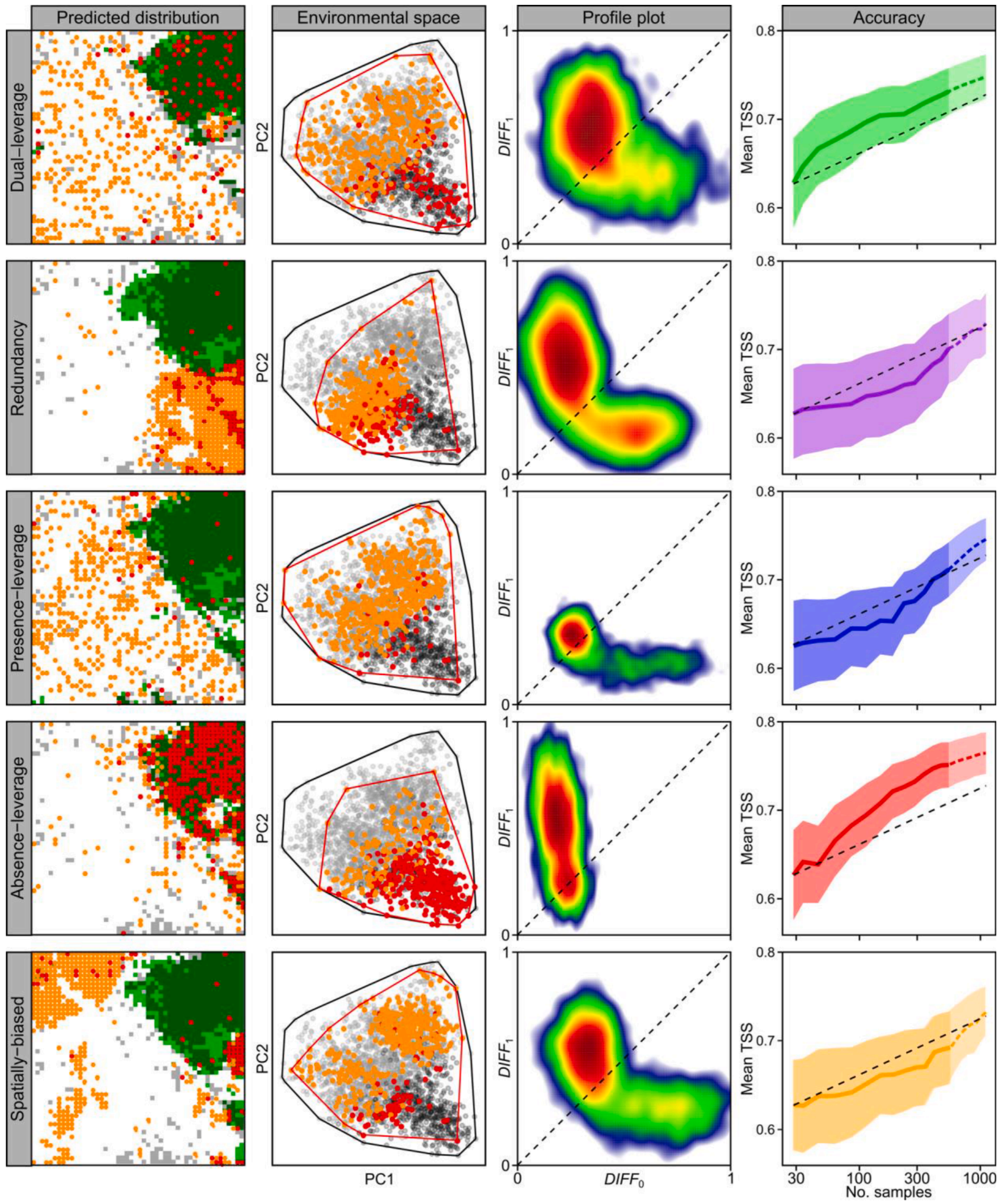

Fig. 6. A snapshot after 13 iterations of accumulating sampling points (480 points) using SDM profiling for a virtual species with an aggregated distribution and medium prevalence (0.25) from Fig. 4. Additional points were selected based upon the four leverage measures from the profile plot: dual-leverage (green), redundancy (purple), presence-leverage (blue), absence-leverage (red), and in a spatially-biased manner independent of the species' distribution (orange). In the 1st column, the predicted distribution (dark green = correctly predicted occurrences; light green = incorrectly predicted occurrences) after 480 points have been selected (orange points: absence; red points: presence) overlaid on to the true distribution (light grey). 2nd column) the environmental space visualised in a PCA biplot for the total area (black polygon) and that sampled by the selected point (red polygon). Grey points are the locations of all cells (light grey: absence; dark grey: presence), and coloured cells those sampled by each strategy (orange: absence; red: presence). 3rd column) A heat map of the profile plot. The density of points goes from dark blue (low density) to dark red (highest density). 4th column). The accuracy measured through the True Skills Statistic (TSS; ± 95% c.i.) as sampling effort is increased in log-space. Coloured dashed lines represent accuracy increases across all iterations. Dashed black lines is the linear increase in accuracy in log-space from the accuracy with 25 sampling points and with 975 sampling points averaged across all species and methods, and serves as a visual aid.

suspicious points in existing datasets (e.g. occurrence records with geocoordinate issues or misidentifications).

4.1. Visualisation of leverage and spatial biases

If we calculate leverage for every cell, we can plot the four leverage distances individually as spatially-explicit maps (Fig. 7). Alternatively, as there are only three degrees of freedom (see 2.2, above), we can visualise all leverages simultaneously by mapping three of the leverage values on to three axes using a red-green-blue (RGB) plot. In the example in Fig. 7, values for absence-leverage determine the values for red, dual-leverage the values for green and blue is determined by presence-

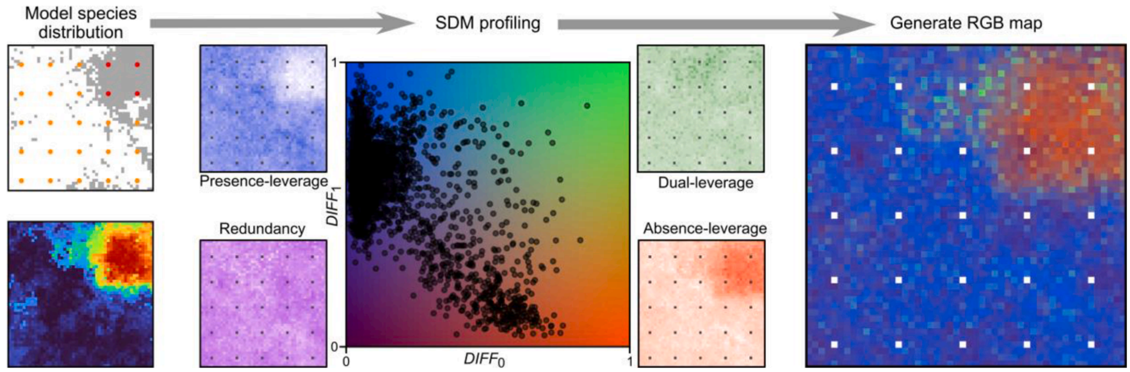

Fig. 7. We can generate spatially-explicit maps for each leverage measure (the four individual maps surrounding the profile plot), or all four measures can be visualised as an RGB plot (in this example red = absence-leverage, green = dual-leverage, and blue = presence-leverage).

leverage. Now areas in blue are likely absences (high presence-leverage), areas in red likely presences (high absence-leverage), areas in purple of little importance for model accuracy (high redundancy), and light green indicates high dual-leverage.

4.2. Evaluating and improving existing sampling schemes

Existing monitoring schemes have limited sampling capacity due to financial and methodological constraints. Therefore producing SDMs for target species can be an important analytical tool in conservation and management where complete sampling of all sites is not feasible (Domisch et al., 2015; Kuemmerlen et al., 2016, 2015). In most cases, the location of sampling sites is based on an equally-spaced, or on an environmentally-stratified design with some degree of randomness within those constraints. All these cases are insensitive to the species perspective of its environment, and to the model's ability to distinguish between presences and absences. Thus, it is important to develop tools that will optimize sampling sites selection for existing or new monitoring schemes (Guisan et al., 2006).

As the simulations show in the previous section, SDM profiling may be used to select new sampling sites that will likely lead to the largest increases in model accuracy for the minimum increase in sampling effort (Figs 5, 6). Interestingly, the simulations showed that, contrary to expectations, selecting new sites so as to cover a broad range of environmental conditions (e.g. an environmentally-stratified design) is generally inefficient when monitoring a single species, although such a scheme may be advantageous for assemblage monitoring where species have differing environmental preferences. Instead concentrating on selecting those sites that are at the environmental boundaries between areas of presence and absence is more likely to result in higher accuracy.

The SDM profiling approach differs fundamentally from SDM targeted approaches (Guisan et al., 2006), that base further sampling only on the output of the SDM using sampled locations. Such approaches, such as adaptive niche-based sampling developed for rare species (Chiffard et al., 2020), typically prioritise detecting new presences by focussing on further sampling in unsampled locations with predicted high probability of occurrences, or targeting areas of environmental space that remain unsampled. In contrast, by assessing model leverage, SDM profiling examines the interactions between unsampled environmental space, the prevalence of that environmental space in the landscape, and the effect sizes of the difference between the unsampled environmental space from sampled environmental space estimated from the response curves recovered from the SDM. Therefore we are prioritising sampling that will increase the accuracy of model predictions rather than necessarily discovering new locations of presences (or absences).

We illustrate this potential using the monitoring data of the freshwater gastropod *Ancylus fluviatilis* in the RMO dataset. The scheme is to be extended to higher order streams and 50 candidate sites were drawn up (Fig. 8). By applying SDM profiling to the candidate sites we can

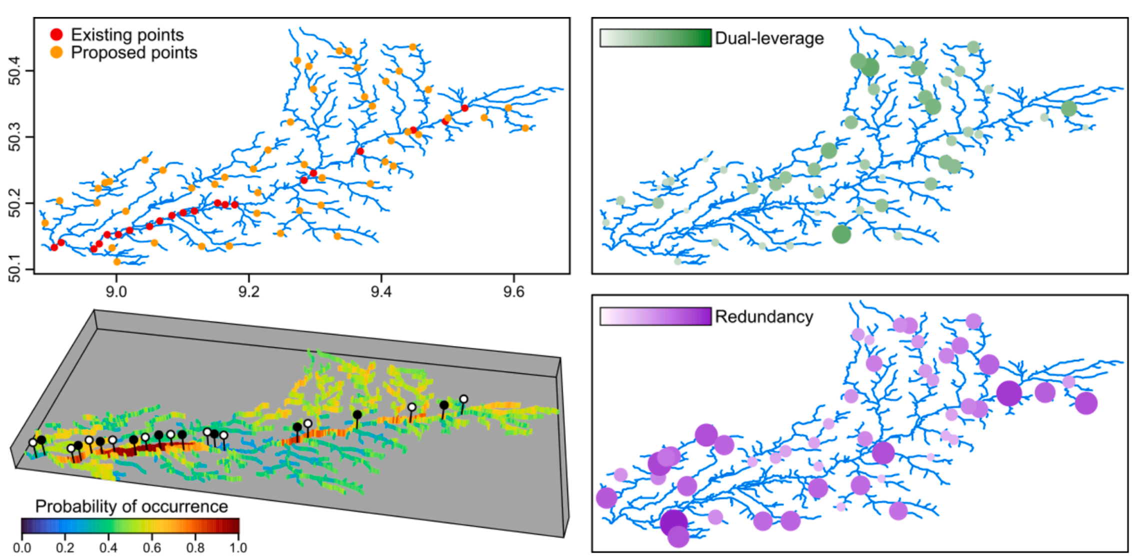

Fig. 8. Selecting new monitoring sites on the LTER site Rhine-Main-Observatory for the freshwater gastropod *Ancylus fluviatilis*. 50 candidate sites were identified (orange points) to supplement 21 existing sites (red points). After SDM profiling on the 50 sites, new sites can be selected as those with the highest dual-leverage values (green), or alternatively we can eliminate candidates with high redundancy-values (purple).

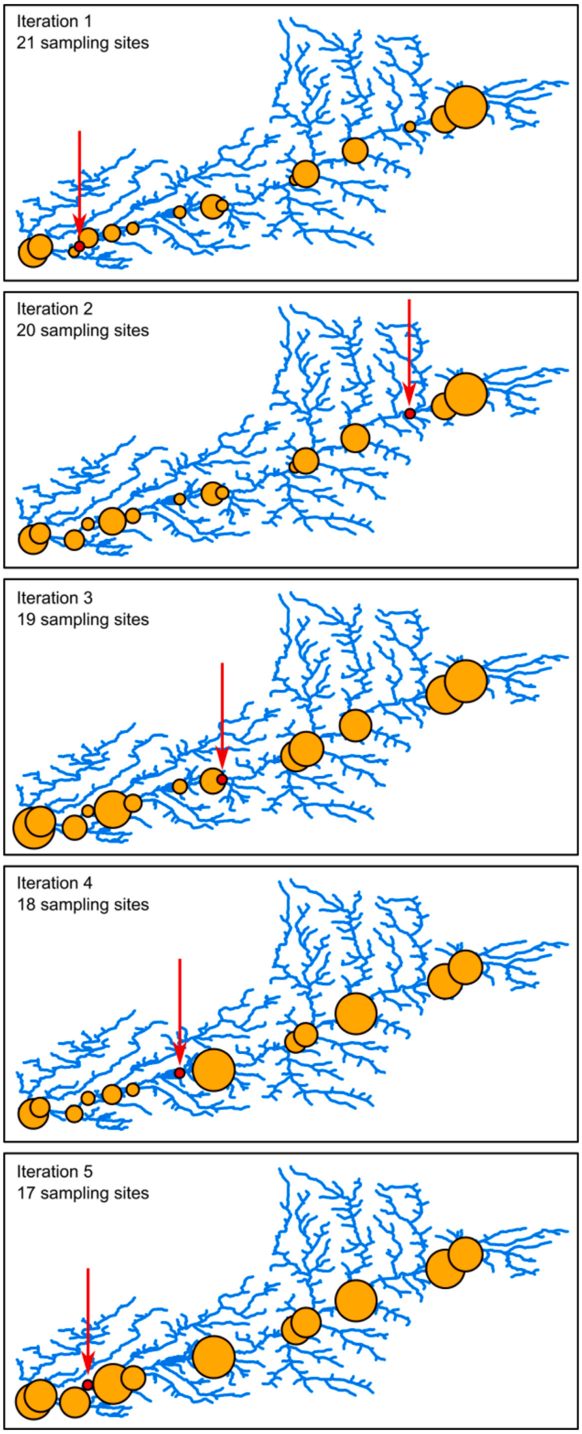

Fig. 9. Example of iterative removal of monitoring sites at the Rhine-Main-Observatory for the freshwater gastropod *Ancylus fluviatilis*. Each site is removed and the deviance measured between the SDMs with and without the site (orange circles). A small deviance indicates little effect of the monitoring site. Five sites are removed sequentially by removing the site with the smallest deviance in each iteration.

select those that have a large, but unknown, potential effect on our knowledge of the species distribution as the sites with the highest dual-leverage values. Alternatively, we can choose to omit sites with highest redundancy values that will provide the least new information.

As well as evaluation of unsampled cells, we can similarly assess the leverage of our sampled sites, for example if we wanted to reallocate effort from existing monitoring sites. In this case we remove each sample

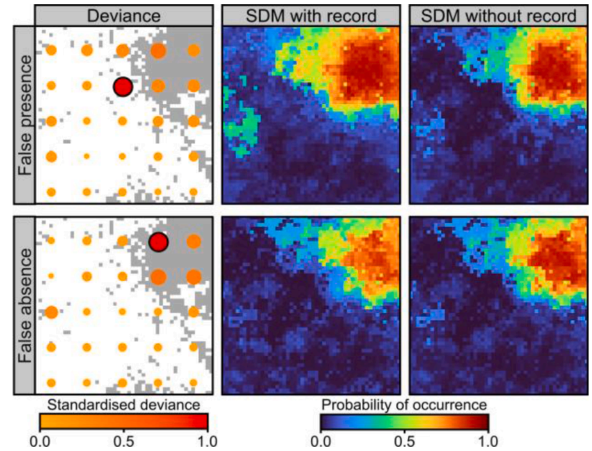

Fig. 10. Flagging of false records using SDM profiling. In each case a false presence (top row) and a false absence (bottom row) were added to the simulated monitoring data (points with black outline). Then we measure the standardised deviance (left column; background shows the true species distribution in grey) between an SDM using all points (centre column) and an SDM removing each point in turn. In both cases removing the false record led to the largest changes in SDM predictions (right column).

point in turn and re-run our SDM, calculating the deviance between the new and base SDMs and standardising each site relative to the largest deviance value. Sites where models are very similar when built with and without that data point (i.e. high redundancy values) are therefore ones with little leverage on our model and so could be removed entirely with small impacts on model accuracy. In the example in Fig. 9 we remove 5 points iteratively by repeating the leverage analysis each time the point with the lowest deviance is removed.

4.3. Flagging suspicious records

Similarly, high deviance after removing a point might indicate incorrect status for that point. This can be a false absence due to, for example, insufficient sampling effort, or alternatively a false presence, which can occur in ad-hoc sampling records, such as that found in mixed-origin datasets such as the Global Biodiversity Information Facility (GBIF; http://www.gbif.org). Any dataset may contain erroneous records, for example due to misidentification, a record of a transient individual, or an error in the assigned coordinates (Maldonado et al., 2015; Yesson et al., 2007), and much effort is now exerted in finding and removing such cases (Chapman et al., 2020; Zizka et al., 2020). A procedure such as that described above for assessing existing monitoring sites, might provide an automated method for flagging up potential errors in datasets, or points that have large influence that can then be expert-assessed.

Returning to the simulated monitoring scheme from Fig. 7 we change one of the absences to a false presence (Fig. 10, top row) and then separately one of the presences to a false absence (Fig. 10, bottom row). Similar to the previous section we then measure the total deviance between an SDM built using all 20 points (including the false record in each case), and then an SDM where each point was removed in turn. In both cases removing the false record led to the largest total deviance in the predicted probabilities between the new and original SDMs.

5. Conclusions

We present a new tool for analysing species distribution models in a spatially-explicit manner, SDM profiling. We take advantage of the fact that an unsampled cell can only be in one of two potential states, presence or absence. We can therefore measure the leverage that a cell has on a SDM through the change in predicted probability of occurrences should that cell be sampled in the future and the species be found to occur there or not. As well as providing a visual aid for identifying areas of high and low certainty in our model predictions, we believe SDM profiling may also provide valuable information for designing and refining monitoring schemes.

Of course, there are constraints on processing time from the number of sampling points and the extent of the modelling area as the number of unsampled cells increases, as well as the complexity of the modelling procedure. For example, SDM profiling of a 100 × 100 cell grid, with 25 sampled points (i.e. two new SDMs will be created for 9975 unsampled cells), 10 environmental variables and using 500 trees to build the random forest models requires 22 mins 40 secs (Intel Core i7–8750H 2.20 GHz and 32GB RAM), but we recommend in such cases to run SDM profiling across multiple cores using the parallelisation option in the provided R package which will reduce the processing time considerably (4 mins 58 secs across 8 cores on the same machine). Of course, constraints on processing time may also be alleviated by only carrying out the profiling for a subset of cells that are of interest (e.g. Fig. 8) or modelling decisions, such as reducing the number of environmental variables, model complexity or algorithm.

Although SDM profiling will inevitably increase the time used to create SDMs, we believe it can provide the spatially-explicit evaluation of model outputs that is currently lacking from the majority of workflows. To maximise the usage of limited funds, we also encourage the consideration of sampling and modelling strategies at the planning stage of any project, as part of an overall strategy to better predict species distribution patterns (Jeliazkov et al., 2022). We therefore hope that SDM profiling can become an important tool to for the adaptive optimisation of monitoring and conservation projects.

CRediT authorship contribution statement

Charles J. Marsh: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yoni Gavish:** Conceptualization, Methodology, Software, Writing – review & editing. **Mathias Kuemmerlen:** Data curation, Investigation, Software, Writing – review & editing. **Stefan Stoll:** Data curation, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Peter Haase:** Data curation, Funding acquisition, Project administration, Supervision, Writing – review & editing. **William E. Kunin:** Conceptualization, Methodology, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

RMO data is publicly available at https://rmo.senckenberg.de/search/home.php. Simulated data can be generated using the R package downloadable from https://github.com/charliem2003/sdmProfiling

Acknowledgements

Funding: This work was financed by the EU BON project (www.eubon.eu) that is a 7th Framework Programme funded by the European Union under Contract No. 308454. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards 2015). PH received additional funding from the EU Horizon 2020 project eLTER PLUS (Grant Agreement No. 871128).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2022.110170.

References

Aizpurua, O., Paquet, J.-.Y., Brotons, L., Titeux, N., 2015. Optimising long-term monitoring projects for species distribution modelling: how atlas data may help. Ecography 38, 29–40. https://doi.org/10.1111/ecog.00749.

Bahn, V., McGill, B.J., 2013. Testing the predictive performance of distribution models. Oikos 122, 321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x.

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Beale, C.M., Lennon, J.J., 2012. Incorporating uncertainty in predictive species distribution modelling. Philos. Trans. Royal Soc. B: Biol. Sci. 367, 247–258. https://doi.org/10.1098/rstb.2011.0178.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Modell. 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012.

Brotons, L., Herrando, S., Pla, M., 2007. Updating bird species distribution at large spatial scales: applications of habitat modelling to data from long-term monitoring programs. Divers. Distrib. 13, 276–288. https://doi.org/10.1111/j.1472-4642.2007.00339.x.

Chapman, A.D., Belbin, L., Zermoglio, P.F., Wieczorek, J., Morris, P.J., Nicholls, M., Rees, E.R., Veiga, A.K., Thompson, A., Saraiva, A.M., James, S.A., Gendreau, C., Benson, A., Schigel, D., 2020. Developing standards for improved data quality and for selecting fit for use biodiversity data. Biodivers. Inform. Sci. Standards.

Charney, N.D., Record, S., Gerstner, B.E., Merow, C., Zarnetske, P.L., Enquist, B.J., 2021. A test of species distribution model transferability across environmental and geographic space for 108 western North American tree species. Front. Ecol. Evol. 9.

Chiffard, J., Marciau, C., Yoccoz, N.G., Mouillot, F., Duchateau, S., Nadeau, I., Fontanilles, P., Besnard, A., 2020. Adaptive niche-based sampling to improve ability to find rare and elusive species: simulations and field tests. Methods Ecol. Evol. 11, 899–909. https://doi.org/10.1111/2041-210X.13399.

Domisch, S., Jähnig, S.C., Simaika, J.P., Kuemmerlen, M., Stoll, S., 2015. Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. Fund. Appl. Limnol. 45–61. https://doi.org/10.1127/fal/2015/0627.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Ewers, R.M., Marsh, C.J., Wearn, O.R., 2010. Making statistics biologically relevant in fragmented landscapes. Trends Ecol. Evol. 25, 699–704. https://doi.org/10.1016/j.tree.2010.09.008.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. https://doi.org/10.1017/S0376892997000088.

Gräler, B., Pebesma, E.J., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R J. 8, 204–218.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecol. Biogeogr. 24, 276–292. https://doi.org/10.1111/geb.12268.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann, N.E., 2006. Using niche-based models to improve the sampling of rare species. Conserv. Biol. 20, 501–511. https://doi.org/10.1111/j.1523-1739.2006.00354.x.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H. P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435. https://doi.org/10.1111/ele.12189.

Hallman, T.A., Robinson, W.D., 2020. Deciphering ecology from statistical artefacts: competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. Divers Distrib. 26, 315–328. https://doi.org/10.1111/ddi.13030.

He, H.S., DeZonia, B.E., Mladenoff, D.J., 2000. An aggregation index (AI) to quantify spatial patterns of landscapes. Landsc. Ecol. 15, 591–601. https://doi.org/10.1023/A:1008102521322.

Jeliazkov, A., Gavish, Y., Marsh, C.J., Geschke, J., Brummitt, N., Rocchini, D., Haase, P., Kunin, W.E., Henle, K., 2022. Sampling and modelling rare species: conceptual

guidelines for the neglected majority. Glob. Chang. Biol. 28, 3754–3777. https://doi.org/10.1111/gcb.16114.
Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379. https://doi.org/10.1111/ddi.12096.
Kremen, C., Cameron, A., Moilanen, A., Phillips, S.J., Thomas, C.D., Beentje, H., Dransfield, J., Fisher, B.L., Glaw, F., Good, T.C., Harper, G.J., Hijmans, R.J., Lees, D. C., Louis, E., Nussbaum, R.A., Raxworthy, C.J., Razafimpahanana, A., Schatz, G.E., Vences, M., Vieites, D.R., Wright, P.C., Zjhra, M.L., 2008. Aligning conservation priorities across taxa in Madagascar with high-resolution planning Tools. Science 320, 222–226. https://doi.org/10.1126/science.1155193.
Kuemmerlen, M., Schmalz, B., Cai, Q., Haase, P., Fohrer, N., Jähnig, S.C., 2015. An attack on two fronts: predicting how changes in land use and climate affect the distribution of stream macroinvertebrates. Biol 60, 1443–1458. https://doi.org/10.1111/fwb.12580.
Kuemmerlen, M., Stoll, S., Sundermann, A., Haase, P., 2016. Long-term monitoring data meet freshwater species distribution models: lessons from an LTER-site. Ecological Indicators, The value of long-term ecosystem research (LTER): Addressing global change ecology using site-based data 65, 122–132. https://doi.org/10.1016/j.ecolind.2015.08.008.
Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M., Bellard, C., 2018. Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. J. Biogeogr. 45, 1994–2002. https://doi.org/10.1111/jbi.13402.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2, 18–22.
Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N., Antonelli, A., 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecol. Biogeogr. 24, 973–984. https://doi.org/10.1111/geb.12326.
T. Mirtl, M., Borer, E., Djukic, I., Forsius, M., Haubold, H., Hugo, W., Jourdan, J., Lindenmayer, D., McDowell, W.H., Muraoka, H., Orenstein, D.E., Pauw, J.C., Peterseil, J., Shibata, H., Wohner, C., Yu, X., Haase, P., 2018. Genesis, goals and achievements of Long-Term Ecological Research at the global scale: a critical review of ILTER and future directions Sci. Total Environ. 626, 1439–1462. https://doi.org/10.1016/j.scitotenv.2017.12.001.
Moerman, D.E., Estabrook, G.F., 2006. The botanist effect: counties with maximal species richness tend to be home to universities and botanists. J. Biogeogr. 33, 1969–1974. https://doi.org/10.1111/j.1365-2699.2006.01549.x.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197. https://doi.org/10.1890/07-2153.1.
Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11, 4540. https://doi.org/10.1038/s41467-020-18321-y.
R Core Team, 2020. R: a language and environment for statistical computing.
Radosavljevic, A., Anderson, R.P., 2014. Making better MAXENT models of species distributions: complexity, overfitting and evaluation. J. Biogeogr. 41, 629–643. https://doi.org/10.1111/jbi.12227.
Riaz, M., Kuemmerlen, M., Wittwer, C., Cocchiararo, B., Khaliq, I., Pfenninger, M., Nowak, C., 2020. Combining environmental DNA and species distribution modeling to evaluate reintroduction success of a freshwater fish. Ecol. Appl. 30, e02034. https://doi.org/10.1002/eap.2034.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881.
Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35, 211–226. https://doi.org/10.1177/0309133311399491.
Swanson, A.K., Dobrowski, S.Z., Finley, A.O., Thorne, J.H., Schwartz, M.K., 2013. Spatial regression methods capture prediction uncertainty in species distribution model projections through time: prediction uncertainty of SDMs. Global Ecol. Biogeogr. 22, 242–251. https://doi.org/10.1111/j.1466-8238.2012.00794.x.
Syfert, M.M., Smith, M.J., Coomes, D.A., 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PLoS ONE 8, e55158. https://doi.org/10.1371/journal.pone.0055158.
Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2021. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol. Monogr. 92 https://doi.org/10.1002/ecm.1486 e01486.
Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., Culham, A., 2007. How global is the Global Biodiversity Information Facility? PLoS ONE 2, e1124. https://doi.org/10.1371/journal.pone.0001124.
Zizka, A., Carvalho, F.A., Calvente, A., Baez-Lizarazo, M.R., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Moreira, F.G.L., Santos, N.M.C., Santos, T.A.B., Santos-Costa, R.C.dos, Serrano, F.C., Silva, A.P.A.da, Soares, A.de S., Souza, P.G.C.de, Tomaz, E.C., Vale, V.F., Vieira, T.L., Antonelli, A., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.