1-27-2023

itsdm: Isolation forest-based presence-only species distribution modelling and explanation in r

Lei Song
Clark University, lsong@clarku.edu

Lyndon Estes
Clark University, lestes@clarku.edu

Received: 13 September 2022 | Accepted: 10 January 2023

DOI: 10.1111/2041-210X.14067

Methods in Ecology and Evolution

APPLICATION

ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R

Lei Song | Lyndon Estes

Graduate School of Geography, Clark University, Worcester, Massachusetts, USA

Correspondence
Lei Song
Email: lsong@clarku.edu

Funding information
Future Investigators in NASA Earth and Space Science and Technology (FINESST), Grant/Award Number: 80NSSC20K1640

Handling Editor: Giovanni Strona

Abstract

1. Multiple statistical algorithms have been used for species distribution modelling (SDM). Due to shortcomings in species occurrence datasets, presence-only methods (such as MaxEnt) have become increasingly widely used. However, sampling bias remains a challenging issue, particularly for density-based approaches. The Isolation Forest (iForest) algorithm is a presence-only method less sensitive to sampling patterns and over-fitting because it fits the model by describing the unsuitable instead of suitable conditions.
2. Here, we present the ITSDM package for species distribution modelling with iForest, which provides a workflow wrapper for the algorithms in iForest family and convenient tools for model diagnostic and post-modelling analysis.
3. ITSDM allows users to fit and evaluate an iForest SDM using presence-only occurrence data. It also helps the users to understand relationships between species and the living environment using Shapley values, a suggested technique in explainable artificial intelligence (xAI). Additionally, ITSDM can make spatial response maps that indicate how species respond to environmental variables across space and detect areas potentially affected by a changing environment.
4. We demonstrated the usage of the ITSDM package and compared iForest with other mainstream SDMs using virtual species. The results enlightened that iForest is an advantageous presence-only SDM when the actual distribution range is unclear.

KEYWORDS
explainable artificial intelligence (xAI), isolation forest, presence only, shapley values, species distribution modelling (SDM)

1 | INTRODUCTION

Statistical methods and associated algorithms have been used for decades to develop species distribution models (SDMs) (Guisan & Zimmermann, 2000) because of their practical usefulness in ecological decision-making and conservation planning. Their usage continues to expand as ever-growing accessibility of occurrence data from public databases (Sofaer et al., 2019). However, many species occurrence datasets were not gathered in structured surveys. They thus may partially cover suitable habitats, contain sampling issues

and often lack absence cases, which negatively impact SDMs (Beck et al., 2014). Background sampling is a common way to deal with the missing absence records, but without enough background knowledge and proper sampling strategies, taking background samples as pseudo-absence records may confound the environmental response of the modelled species (Barbet-Massin et al., 2012). Anomaly detection algorithms, which are semi-supervised, can take presence-only records without any pseudo-absence samples, reducing the risk of adding false information. Among them, maximum entropy (MaxEnt) has risen to dominance and has been widely used in many case studies (Elith et al., 2011; Phillips & Dudík, 2008). However, as with other density-based methods, it is easily affected by sampling patterns and overfits (Kramer-Schadt et al., 2013; Merow et al., 2013; Radosavljevic & Anderson, 2014). The regularization multiplier may reduce the issues by controlling model complexity, but the effects are species-specific and prone to sample size (Morales et al., 2017; Radosavljevic & Anderson, 2014). Isolation Forest (iForest) uses a novel approach based on the depth of the branch in the tree to calculate the probabilities (Liu et al., 2008, 2010, 2012). Optimizing to unsuitable conditions and not relying on sample density to fit the model, it thus suffers less from overfitting and sampling issues. The tree structure functions similar to profile models that describe the species–environment relationship as an 'environmental profile' (Franklin, 2010), and consequently trends to predict the environmental suitability rather than the probability of detection.

In computer science, iForest is widely applied in spatial and non-spatial anomaly detection and one-class classification problems (Feremans et al., 2020; Khan et al., 2019; Li et al., 2019), but it has not yet been adopted widely in ecology relatively because iForest lacks a standard toolkit to assess ecological validity and produce detailed summaries of fitted relationships. Many post hoc methods have been proposed to analyse the behaviours of models with non-interpretable structures. The 'evaluation strip' technique (Elith et al., 2005) can visualize variable responses for any modelling approach, aiding the users in evaluating and comparing models with different structures. Phillips et al. (2006) implemented leave out one Jackknife test to identify variables with significant individual effects. Shapley values technique is listed as one of the post hoc model agnostic tools in explainable artificial intelligence (xAI) and is encouraged to be applied in SDM research domain (Ryo et al., 2021). It can explain the relative contribution of each feature to the prediction at a given instance locally and summarize variable response and variable importance globally (Lundberg & Lee, 2017; Shapley, 1953).

To take advantage of the ability of iForest to handle presence-only data with less sensitivity to sampling patterns, we developed a new R (R Core Team, 2021) package ITSDM. It provided a wrapper for iForest and its related variants (Cortes, 2021b, 2022; Guha et al., 2016; Hariri et al., 2019; Liu et al., 2008, 2010) to do species distribution modelling, alongside methods delivering ecological insights from the model. This package aims to provide ecological modellers with an additional tool for creating SDMs, which can complement well-established existing approaches, such as those implemented in BIOMOD (Thuiller et al., 2009, 2021).

2 | PACKAGE STRUCTURE AND DESCRIPTION

ITSDM is a workflow wrapper coded in R and knits iForest and Shapley value explanation into an SDM workflow. The package's functions are in four groups (Table 1): pre-modelling analysis, modelling, model explanation and post-modelling analysis. The pre-modelling analysis functions diagnose the relationships between environmental variables and target potential sampling errors in the occurrence dataset. The model implementation functions format observation dataset, build and evaluate the model with different user settings. The model explanation functions delineate the importance of environmental variables and the species' spatial and non-spatial responses to them. The package also contains a post-modelling toolkit for further analysis of modelling results, for instance, analysing the impacts of a changing environmental variable, converting predicted suitability to a presence–absence and comparing the contribution of environmental variables to observations. Importantly, all Shapley value-based functions (such as *shap_dependence*, *shap_spatial_response*, *detect_envi_change* and *variable_contrib*) can apply to any fitted models as long as the function inputs are correctly set (see example in Section 4.2). Because visualization is critical in ecological modelling, ITSDM provides corresponding *print* and/or *plot* generic functions to visualize every object (Table 1).

3 | SDMs WITH ISOLATION FOREST

Isolation Forest (iForest) is built based on the decision tree architecture to distinguish anomalies or outliers from a set of samples (e.g. presence-only samples). Because the majority of the samples are normal, anomalies are few and different. In presence-only SDM, it means samples gathered in less suitable areas are lower in quantity and environmentally different from samples in suitable areas. iForest aims to fit a model to describe these anomalies rather than the normal samples, therefore, does not necessarily need background samples. More importantly, it is more robust to sampling issues and overfitting.

iForest uses the path in the tree structure to calculate the probability of a sample being anomalous (termed as anomaly score). Reflecting on a tree structure, the anomalies are isolated closer to the tree's root note, so they have shorter paths (Figure 1a). Given an isolation tree built on a dataset $X = \{x_1, \ldots, x_n\}$ of n samples, X is divided recursively by a test for every internal node T_{in} to two sibling nodes (T_l, T_r) until the node becomes an external node T_{ex} with no child or a predefined depth limit is reached. Within the feature space, the test is a hyperplane defined by a random normal vector and intercept (Figure 1b). The node splitting criterion for a given point $\vec{x}$ is as follows (Hariri et al., 2019):

$$(\vec{x} - \vec{p}) \cdot \vec{n} \leq 0, \quad (1)$$

where $\vec{n}$ is a random normal vector uniformly over the unit N-Sphere which specifies the $N-1$ dimensional hyperplane to split

TABLE 1 Core functions and descriptions in ITSDM

Function (object)		Visualization	Description
Pre-modelling analysis	dim_reduce (ReducedImageStack)	print	Select numeric environmental variables with pairwise Pearson correlation lower than a defined threshold. The user can specify preferred variables
	suspicious_env_outliers (EnvironmentalOutlier)	print, plot	Detect suspicious environmental outliers in the occurrence dataset according to each environmental covariate's general condition
Modelling	format_observation (FormatOccurrence)	print	Quickly format the dataset to fit into the ITSDM workflow
	isotree_po (POIsotree)	print	Build Isolation Forest-based SDM and do the related model explanation, which optionally calls model explanation functions
	evaluate_po (POEvaluation)	print, plot	Evaluate the model based on presence-only data
Model explanation	variable_analysis (VariableAnalysis)	print, plot	Evaluate environmental variable importance using leave one out Jackknife test and Shapley values
	marginal_response (MarginalResponse)	plot	Calculate marginal response curve of environmental variables using the evaluation strip method proposed by Elith et al. (2005)
	independent_response (IndependentResponse)	plot	Calculate independent response curve of environmental variables by creating an independent model using only one variable each time (Phillips et al., 2006)
	shap_dependence (ShapDependence)	plot	Calculate Shapley value-based variable dependence plot which is introduced in Section 4.2. It is a supplementary plot for marginal and independent response curves using a completely different method. It also can explore the relationship between two environmental variables
	spatial_response (SpatialResponse)	plot	Generate spatially partial dependence maps with type of marginal, independent and Shapley value based
	shap_spatial_response (SHAPSpatial)	plot	Generate spatially partial dependence maps only with Shapley values. It can work on other external models
Post-modelling analysis	detect_envi_change (EnviChange)	print, plot	Use post hoc Shapley values technique to detect the tipping points and potentially affected areas due to a changing environmental variable
	convert_to_pa (PAConversion)	print, plot	Convert predicted suitability to presence–absence map
	variable_contrib (VariableContribution)	print, plot	Evaluate local and/or global variable contributions for interested observation(s)

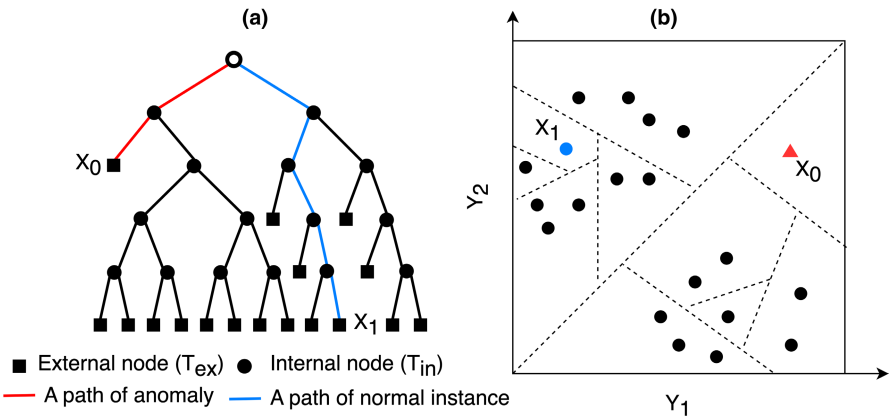

FIGURE 1 Schematic representation of a single tree (a) and its feature space (b) for an Extended Isolation Forest (EIF) built by a two-dimensional dataset.

nodes for a dataset with N attributes (N-dimensional). $\vec{p}$ is a set of values from a uniform distribution over the range of possible values at each node which serves as a set of random intercepts of the hyperplane. This is a general definition of iForest called Extended Isolation Forest (EIF). Standard iForest is a special case of EIF whose split test only consists of a randomly selected attribute q

and a split value p such that the test $q < p$ splits the node into sibling nodes (Liu et al., 2008, 2012).

After the whole dataset is split into trees, an anomaly score is calculated as follows:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}, \tag{2}$$

where $h(x)$ is the path length for external node terminations in one tree, and $E(h(x))$ is the mean $h(x)$ of all trees. $c(n)$ is the normalizing factor (Preiss, 2008). The calculated anomaly score s ranges from 0 to 1: the closer to 1, the more likely the sample is anomalous; otherwise, the sample is more likely to be normal. To fit iForest into the SDM workflow, ITSDM uses a linear conversion ($p = 1 - s$) to translate the anomaly score (s) into environmental suitability (p).

As a popular algorithm in anomaly detection, iForest has been continuously improved with amended node splitting methods such as split-criterion iForest (Liu et al., 2010), Robust random cut forest (Guha et al., 2016) and Fair-cut forest (Cortes, 2021b), as well as new metrics for calculating outlier scores (Cortes, 2021a). The R (R Core Team, 2021) package ISOTREE (Cortes, 2022) is an ensemble of iForest and these variants with fast and multi-threaded implementation and thus is used in ITSDM for model training. Table S1-1 (Appendix S1) lists the decisive arguments used in the function *isotree_po* (Table 1) for specific model types in the family of iForest.

4 | APPLICATION OF SHAPLEY VALUES

4.1 | Local explanation and applications in ITSDM

The Shapley value (Shapley, 1953) is an idea from cooperative game theory, which fairly distributes a game's payouts among players. The SHapley Additive exPlanations (SHAP) is an additive feature attribution method based on Shapley values that decompose individual predictions of a model into the sum of the contributions of each variable value (Lundberg & Lee, 2017). Assume there is a prediction $f(x)$ for a single input x, the additive feature attribution method specifies the explanation as (Lundberg & Lee, 2017):

$$g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x'_i, \tag{3}$$

where g is the explanation model. x' is the simplified x that maps to the original x by function $x = h_x(x')$. M is the number of input features. $\emptyset_0$ is the constant value when all inputs are missing, and $\emptyset_i \in \mathbb{R}$ is the feature attribution for feature i. It was theoretically proved that Shapley values are the unique solution of Equation (3) with three desirable properties (see details in Lundberg & Lee, 2017):

$$\emptyset_i = \sum_{Q \subseteq S \setminus \{i\}} \frac{|Q|!(|S| - |Q| - 1)!}{|S|!} \left[f_{Q\cup\{i\}}\left(x_{Q\cup\{i\}}\right) - f_Q\left(x_Q\right) \right], \tag{4}$$

where S is the set of all features in the model. Q is a subset of S. $f_{Q\cup\{i\}}$ is a model trained with feature i present and f_Q is a model trained with feature i withheld. Thus, $f_{Q\cup\{i\}}\left(x_{Q\cup\{i\}}\right) - f_Q\left(x_Q\right)$ represents the effects of including feature i on the model. Because the effect of

withholding a feature relies on other features, $\emptyset_i$ calculates the weighted average of $f_{Q\cup\{i\}}\left(x_{Q\cup\{i\}}\right) - f_Q\left(x_Q\right)$ of all possible subsets $Q \subseteq S \setminus \{i\}$.Several approaches (e.g. Kernel SHAP and Linear SHAP) (Molnar, 2020; Štrumbelj & Kononenko, 2014) have been proposed to approximate Shapley values (Equation 4). The package FASTSHAP (Greenwell, 2021) is used in ITSDM to estimate Shapley values, in which a Monte-Carlo sampling approach (Štrumbelj & Kononenko, 2014) is efficiently implemented.

Shapely values demonstrate how each explanatory covariate pushes the model result from the base value (the average model output over the training dataset) (Molnar, 2020). Positive values vote for presence, and negative values vote for absence. The higher the absolute Shapley value is, the more important the explanatory variable is. Using Shapley values and the characteristics, function *variable_contrib* (Table 1) in ITSDM can diagnose how the explanatory variables decide the environmental suitability at each observation location.

4.2 | Global explanation and applications in ITSDM

Additionally, Shapley values can be integrated into global explanations such as variable importance and response curves. Because features with large absolute Shapley values are important, variable importance could be evaluated by averaging the absolute Shapley values per feature across the whole data:

$$I_i = \frac{\sum_{j=1}^{n} \left| \emptyset_i^{(j)} \right|}{n}. \tag{5}$$

This is implemented in function *variable_analysis* (Table 1) in package ITSDM.

Shapley values technique shows how a species responds to an environmental variable by plotting all possible feature values $\left\{ x_i^{(j)} \right\}_{j=1}^{n}$ against the corresponding Shapley values $\left\{ \emptyset_i^{(j)} \right\}_{j=1}^{n}$. As Shapley values are signed, the response curves also can show the tipping point(s) of when this species starts to be negatively impacted by this environmental variable. In ITSDM, *shap_dependence* (Table 1) is the function to generate Shapley value-based response curves. To illustrate how a variable affects prediction spatially, ITSDM provides the function *shap_spatial_response* (Table 1), which uses Shapley values to generate spatial response maps.

Expanding from *shap_dependence* and *shap_spatial_response*, ITSDM provides a unique function (*detect_envi_change*) to analyse the vulnerable areas potentially impacted by the changing environmental variables. The users can apply a number to the current environmental variable or assign a completely new future

environmental variable. As a model agnostic post hoc method (Ryo et al., 2021), Shapley values technique can be used to explain any predictive models; therefore, the Shapley value-based functions in ITSDM including *detect_envi_change* can apply to any SDMs. For instance, we fitted a MaxEnt SDM (Phillips & Dudík, 2008) named *mod_maxent* with multiple Bioclimatic variables BIO1, BIO2, BIO3, BIO13, BIO14, BIO18 and BIO19 (Fick & Hijmans, 2017) to estimate the habitat suitability of Za Baobab tree (*Adansonia za Baill.*) in Madagascar (see 'Data availability' section for code). With a *stars* (Pebesma, 2022) object called *bios_current* to represent the current environment, a *stars* (Pebesma, 2022) object called *bios_future* to represent the future (2041–2060) environment (Fick & Hijmans, 2017), and a wrapper function called *pfun* for *mod_maxent* to do prediction, *detect_envi_change* works as follows to detect potential impacts to Za Baobab tree by a changing BIO1 (annual mean temperature):

```
pfun <- function(X.model, newdata) {
 predict(X.model, newdata,
 args = c("outputformat=cloglog"))}
bio1_changes <- detect_envi_change(
 model = mod_maxent,
 var_occ = training[, 2:ncol(training)],
 variables = bios_current,
 target_var = "bio1",
 variables_future = bios_future,
 pfun = pfun)
```

The function returns a response curve with detected tipping points (Figure 2a), a vector of detected tipping points, a map of contribution change (Figure 2b) and a *stars* (Pebesma, 2022) object of contribution change.

5 | EXAMPLE

To demonstrate the package functionality, we provide a short example using a virtual species generated by the package VIRTUALSPECIES (Leroy et al., 2016), the distribution of which is in mainland Africa and shaped by climatic variables bio1, bio5 and bio12 (Fick & Hijmans, 2017). For this example, we took 2000 random presence-only samples and selected bio1, bio5, bio12 and three other unrelated features (var1 through var3) as the explanatory environmental variables. In the workflow, 70% of the samples are used for training, and 30% of them are used for evaluation. The details of the virtual species can be found in Sections 2.1 and 2.2 in Appendix S1.

In function *isotree_po*, a model is fit to the provided *sf* (Pebesma, 2018) object of occurrence points and corresponding environmental variables, along with an optional *sf* (Pebesma, 2018) object of occurrence points for independent evaluation. For example, with a training set of occurrence points *obs*, an independent evaluation set called *eval* and a *stars* (Pebesma, 2022) object holding the environmental predictors (*env_vars*), the following workflow creates an EIF model with an extension level of 2 (see more options in Table S1-1) and a sampling rate of 0.8:

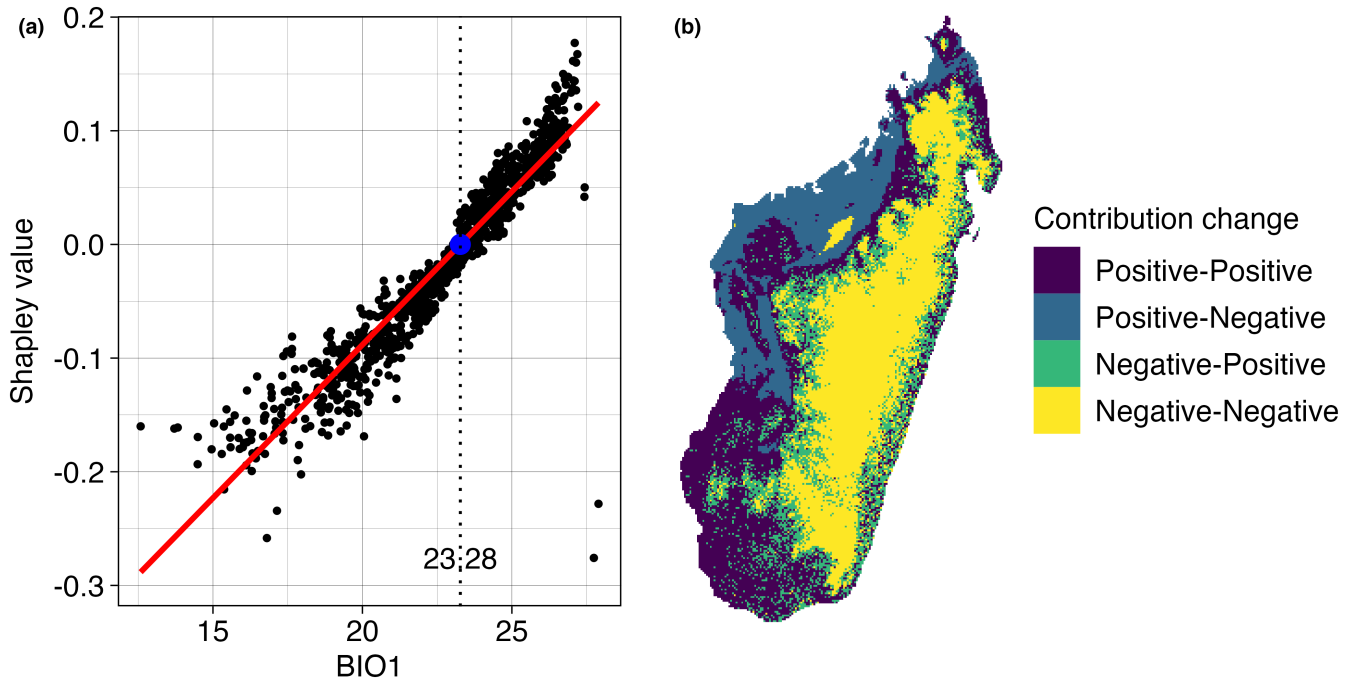

FIGURE 2 Environmental change analysis of BIO1 (annual mean temperature) to Za Baobab tree (*Adansonia za Baill.*) in Madagascar. Panel (a) shows that Za Baobab tree has a positive linear response to annual mean temperature in Madagascar. 23.28°C is the tipping point of annual mean temperature, which means the Za Baobab tree in areas with an annual mean temperature near 23.28°C is vulnerable to a cooling temperature. Panel (b) shows Za Baobab tree in most areas of Madagascar will not be affected by a changing annual mean temperature. The annual mean temperature in Northwest coastal areas will become not suitable for Za Baobab tree.

```
# Create an Extended isolation forest
mod <- isotree_po(
 obs = obs,
 obs_ind_eval = eval,
 variables = env_vars,
 sample_size = 0.8,
 ndim = 2)
```

The function *isotree_po* provides a highly automatic workflow that contains model creation, model evaluation, model prediction and model explanation, with corresponding *print* and/or *plot* options to check the results (Table 1). The full description of the results can be found in Section 2 of Appendix S1. Here, we only present the Shapley value-based analysis.

If argument *check_variable* is set to *FALSE* in function *isotree_po*, the users can call function *variable_analysis* to diagnose variable importance:

```
var_analysis <- variable_analysis(
 model = mod$model,
 pts_occ = mod$observation,
 pts_occ_test = mod$independent_test,
 variables = mod$variables)
plot(var_analysis)
```

The function ranks environmental variables based on Shapley values (Figure 3) as well as the leave-one-out Jackknife test (Section 2.5.1 and Figure S1-5 in Appendix S1).

ITSDM employs several methods to generate response curves, including spatial ones. The Shapley value-based response curve conveys how prediction is pushed away from the average prediction across the whole training dataset (Section 4.2). The Shapley values also allow users to diagnose the correlation between two variables. For example, Shapley value-based response curves of bio1 and bio12 are plotted and coloured by bio5 (Figure 4):

```
# Plot Shapley value-based response curves without smoothing
plot(mod$shap_dependences,
 target_var = c('bio1', 'bio12'),
 related_var = 'bio5', smooth_line = FALSE)
```

It is recommended to use response curves together with variable importance analysis to explain model inputs. However, the standard response curves only provide a graphical, non-spatial assessment of how a variable influences prediction. To illustrate how a variable affects prediction spatially, ITSDM provides the function *spatial_response*, which generates spatial response maps. To calculate response maps, *spatial_response* is used with a non-zero *shap_nsim*:

```
# Make spatial response maps with all three methods
# Make sure to set a non-zero shap_nsim
full_spatial_responses <- spatial_response(
 model = mod$model,
 var_occ = mod$vars_train,
 variables = mod$variables,
 shap_nsim = 10)
plot(full_spatial_responses, target_var = 'bio12')
```

The last line displays the spatial response maps of variable bio12, and the Shapley value-based one is shown in Figure 5. Areas with

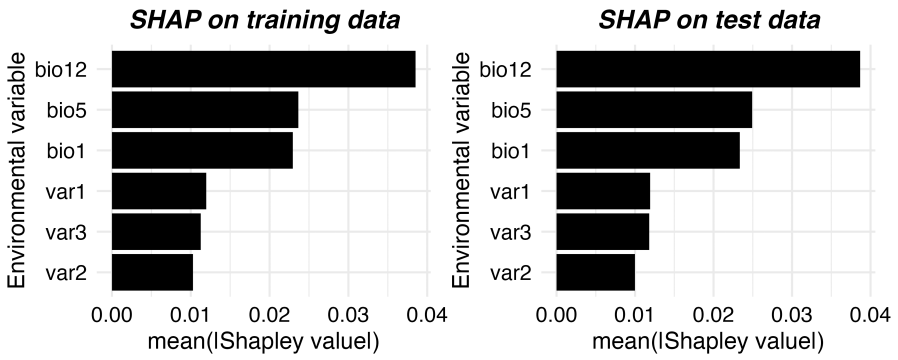

FIGURE 3 Variable importance of virtual species case diagnosed by Shapley values technique. Variables bio12, bio5 and bio1 have much higher importance than var 1 through var 3, as intended. In addition, the similarity in the values for these metrics for both the training and test dataset indicates that the model is generalizable.

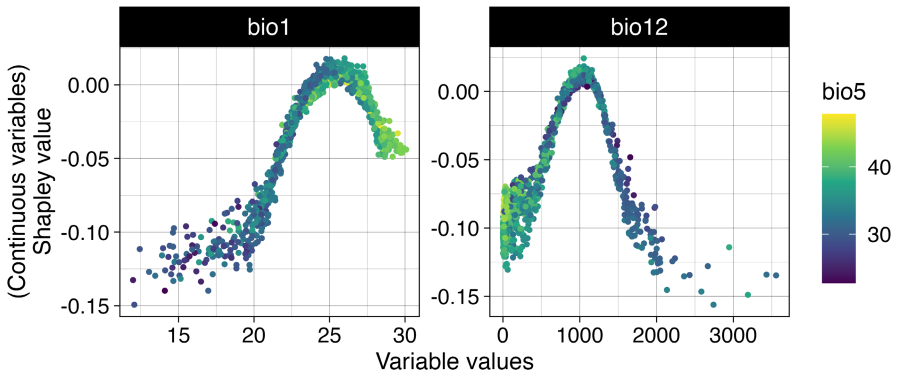

FIGURE 4 Shapley value-based response curves of bio1 and bio12 coloured by bio5 in our virtual species case. The modelled species has a strong positive response to both bio1 and bio12 that, respectively, peak at 25°C and 1000 mm, and that the two are also strongly correlated with bio5, particularly in the upper range for bio1 and in the lower to mid-range for bio12.

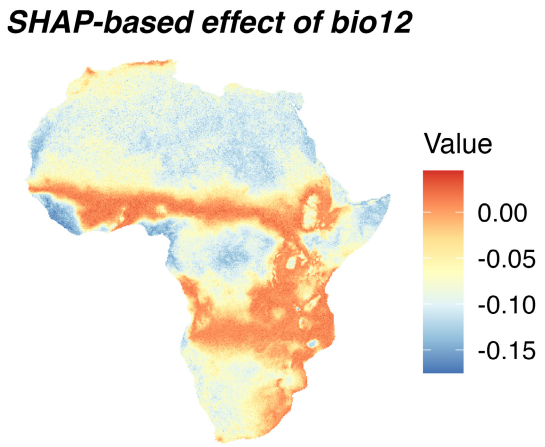

FIGURE 5 Shapley value-based spatial response map of variable bio12 in our virtual species case. It is evident that bio12 contributes minimally in some areas even though it is the most vital environmental variable diagnosed in variable analysis.

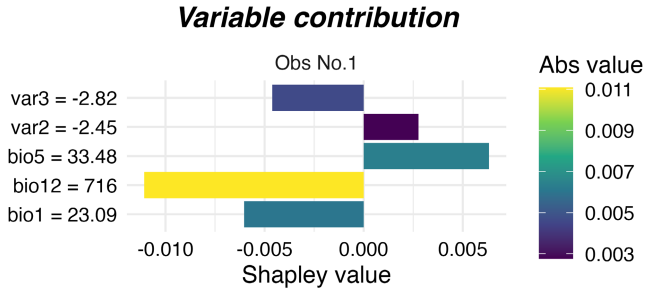

FIGURE 6 Variable contributions to the modelled suitability of an occurrence observation.

Shapley values below zero are where bio12 votes for absence for this species, and areas with Shapley values above zero are the opposite. Variables with large absolute Shapley values contribute more than others (Section 4.2).

ITSDM also includes several optional post-analysis steps (Table 1), such as analysing variable contributions to a specific observation, as shown in Figure 6. The figure shows that bio5 pushes the predicted suitability higher and votes for presence. The bio12 and bio 1 contribute oppositely.

The full list of functions and additional examples can be found in the ITSDM package documentation and package VIGNETTES with extended examples.

6 | COMPARISON WITH OTHER SDMs AND RECOMMENDATIONS

To compare the predictive performance of iForest with other SDMs and highlight the conditions when it is beneficial to use iForest, we generated 50 virtual species with package VIRTUALSPECIES (Leroy et al., 2016) (see 'Data availability' section for code). Bioclimatic variables BIO1, BIO2, BIO5, BIO6, BIO12 and BIO15 (Fick & Hijmans, 2017) in mainland Africa were used to simulate these species. We generated a species suitability map by applying a Gaussian, linear, logistic or quadratic function with random parameters on randomly selected three to five variables for each species (Leroy et al., 2016). The final suitability is a multiplicative function of responses to the selected variables. A threshold of 0.5 or 0.6 was used to convert suitability to presence–absence to represent the normal detection type (Figure 7). A threshold of 0.8 or 0.9 was used to represent the core area concentrated detection type (Figure 7). A prevalence-weighted random number from 100 to 500 of presence-only samples was drawn from presence–absence map for both detection types. In addition, 10,000 background samples were taken for all SDMs except iForest. For evaluation, 2000 presence–absence samples were drawn for each species by excluding all training presence locations and their 3×3 neighbours and then subset the majority class to ensure class balance.

True skill statistics with a threshold of 0.5 ($TSS_{0.5}$) and three threshold-independent evaluation metrics: Area under the ROC curve (AUC), Pearson correlation (COR) and Euclidean distance were used to assess predictive performance. AUC and $TSS_{0.5}$ measure the capability of a model to separate presences from absences. COR values and Euclidean distances in this experiment were calculated between the predicted environmental suitability and the simulated suitability of the virtual species. They work together to measure the similarity between predicted and actual suitability values, which is to say, they have the same values for the same cases.

We selected seven SDMs with high performance (Valavi et al., 2022) to make the comparison: Generalized linear model (GLM), generalized additive model (GAM), maximum entropy (MaxEnt), Random forest (RF), multivariate adaptive regression spline (MARS), boosted regress trees (BRT) and extreme gradient boosting (XGBoost). The results are shown in Figure 7. If the training samples can represent the actual distribution well (Normal case in Figure 7), GAM and GLM perform better than iForest and others, having a greater ability to discriminate presences and absences (high AUC and $TSS_{0.5}$) and higher similarity to actual suitability (high COR and Euclidean distance). It is worth mentioning that suitability values predicted by iForest have the closest Euclidean distance with actual suitability values, which is also evident in Figure S1. If the training samples only represent the actual distribution partially, for example, having sampling bias or imperfect detection (Core area case in Figure 7), iForest starts to be advantageous. Even though iForest gets slightly lower AUC than GLM, GAM and MaxEnt, it makes significantly higher COR and $TSS_{0.5}$ and lower Euclidean distance. The similar performance of models fitted under two cases (Figure 7; Figure S1-1) indicates that iForest is resistant to sampling issues and overfitting.

For presence-only species distribution modelling, when the species occurrences cover the true distribution range, models like GLM, GAM or MaxEnt perform better than iForest. This is also true if occurrences do not cover the whole distribution range, but the range is known so that background samples can be extracted conditionally. When it is unfeasible to estimate the distribution range before modelling, iForest can be a cautious choice.

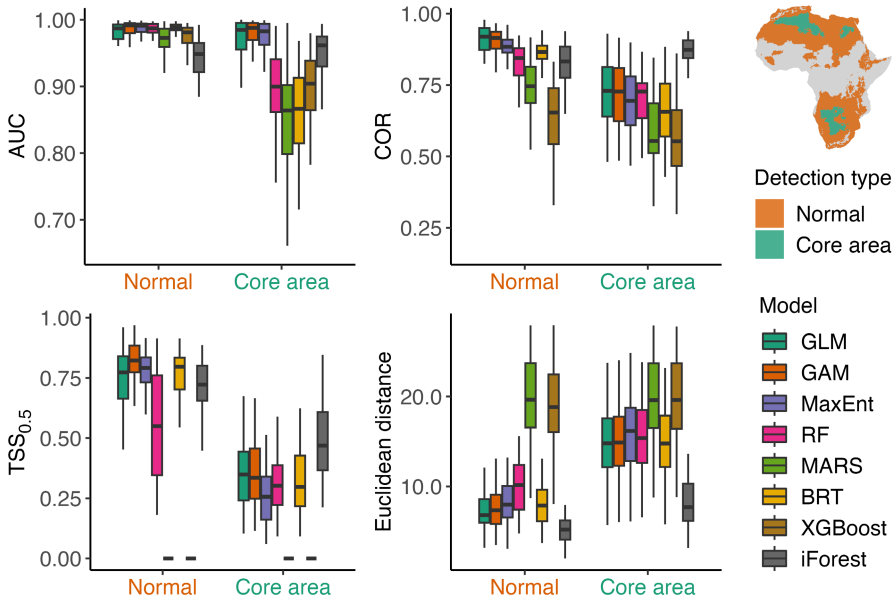

FIGURE 7 Performance comparison between Isolation Forest (iForest) and other mainstream SDM models. Evaluation metrics are area under the ROC curve (AUC), Pearson correlation between modelled and real habitat suitability (COR), True Skill Statistic with a threshold of 0.5 ($TSS_{0.5}$) and Euclidean distance between modelled and real suitability. Normal detection type is the case that the species can be detected in any areas with suitability higher than 0.5 or 0.6 and Core area type is the case that the species can only be detected in areas with suitability higher than 0.8 or 0.9. The figure in upright is an example drawn with No. 6 virtual species.

7 | DISCUSSION

iForest is an appealing method in SDM because it takes presence-only data as input, which matches it with most occurrence datasets of wildlife nowadays. Additionally, splitting feature space by hyperplanes is similar to profile models that translate species–environment relationships into profiles. Thus, it results in environmental suitability rather than the probability of presence. Unlike methods that are optimized to suitable conditions, iForest is optimized to describe unsuitable conditions and thus is less likely to overfit (Abe et al., 2006; He et al., 2003; Rousseeuw & van Driessen, 1999). These give iForest strengths as an SDM, particularly when it is unclear if the presence samples cover suitable areas fully and there are no reliable absence samples to use.

Shapley values technique is a growing topic of interest in interpretable machine learning, as they can help to explain any predictive model (Ryo et al., 2021). It offers a potentially powerful tool to comparably interpret SDMs that are built with different methods and decipher complex models to explain real-world ecological phenomena (Mammola et al., 2019). As a post hoc technique, Shapley values can be used to interpret the impacts of a changing environment in species distribution conveniently.

The R package ITSDM offers convenient functions to fit iForest SDM and generalizes the Shapley values technique for all SDMs to analyse species' response to the environment. Undoubtedly, not relying on causal mechanisms, iForest and Shapley values' technique has the same limitations as other statistical methods in applications of ecological modelling, especially for change analysis. ITSDM is intended as a new SDM toolbox that complements existing frameworks, which will enable users to apply iForest and Shapley values' technique in their studies and explore advantages and disadvantages.

AUTHOR CONTRIBUTIONS

Lei Song conceived the ideas, collected example data and analysed the data; Lei Song and Lyndon D. Estes designed methodology and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This project was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) program (award number: 80NSSC20K1640). The authors thank David Cortes for the suggestion of improving the code flexibility and the authors for all the fabulous and valuable packages that ITSDM depends on.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14067.

DATA AVAILABILITY STATEMENT

The ITSDM package, documentation and example data are hosted and available on CRAN (https://cran.r-project.org/package=itsdm). The source code can be assessed at GitHub (https://github.com/LLeiSong/itsdm), and version 0.2.0 of the package used for this manuscript is archived on Zenodo (Song & Estes, 2023). All simulation species and scripts not shown in supplementary materials are available via Open Science Framework (OSF): https://osf.io/8mc4e/.

ORCID

Lei Song https://orcid.org/0000-0002-4371-1473
Lyndon Estes https://orcid.org/0000-0002-9358-816X

REFERENCES

Abe, N., Zadrozny, B., & Langford, J. (2006). Outlier detection by active learning. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 504–509. https://doi.org/10.1145/1150402.1150459

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

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

Cortes, D. (2021a). Isolation forests: Looking beyond tree depth. *ArXiv:2111.11639 [Cs, Stat]*. http://arxiv.org/abs/2111.11639

Cortes, D. (2021b). Revisiting randomized choices in isolation forests. *ArXiv:2110.13402 [Cs, Stat]*. http://arxiv.org/abs/2110.13402

Cortes, D. (2022). *Isotree: Isolation-based outlier detection*. https://CRAN.R-project.org/package=isotree

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289. https://doi.org/10.1016/j.ecolmodel.2004.12.007

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Feremans, L., Vercruyssen, V., Cule, B., Meert, W., & Goethals, B. (2020). Pattern-based anomaly detection in mixed-type time series. *Machine learning and knowledge discovery in databases*, pp. 240–256.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Franklin, J. (2010). *Mapping species distributions: Spatial inference and prediction*. Cambridge University Press.

Greenwell, B. (2021). *Fastshap: Fast approximate Shapley values*. https://CRAN.R-project.org/package=fastshap

Guha, S., Mishra, N., Roy, G., & Schrijvers, O. (2016). Robust random cut forest based anomaly detection on streams. *International conference on machine learning*, pp. 2712–2721.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation Forest. *IEEE Transactions on Knowledge and Data Engineering*, *1*, 1479–1489. https://doi.org/10.1109/TKDE.2019.2947676

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, *24*(9–10), 1641–1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Khan, S., Liew, C. F., Yairi, T., & McWilliam, R. (2019). Unsupervised anomaly detection in unmanned aerial vehicles. *Applied Soft Computing*, *83*, 105650.

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., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Li, S., Zhang, K., Duan, P., & Kang, X. (2019). Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(1), 319–329.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE international conference on data mining*, pp. 413–422.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2010). On detecting clustered anomalies using SCiForest. *Joint European conference on machine learning and knowledge discovery in databases*, pp. 274–290.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. https://doi.org/10.1145/2133360.2133363

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.

Mammola, S., Milano, F., Vignal, M., Andrieu, J., & Isaia, M. (2019). Associations between habitat quality, body size and reproductive fitness in the alpine endemic spider *Vesubia jugorum*. *Global Ecology and Biogeography*, *28*(9), 1325–1335. https://doi.org/10.1111/geb.12935

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, *5*, e3093. https://doi.org/10.7717/peerj.3093

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. (2022). *Stars: Spatiotemporal arrays, raster and vector data cubes*. https://CRAN.R-project.org/package=stars

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Preiss, B. R. (2008). *Data structures and algorithms with object-oriented design patterns in C++*. John Wiley & Sons.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radosavljevic, A., & Anderson, R. P. (2014). Making better MAXENT models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Rousseeuw, P. J., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212. https://doi.org/10.2307/1270566

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the theory of games* (2.28, pp. 307–317).

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., Edwards, T. C., Guala, G. F., Howard, T. G., Morisette, J. T., & Hamilton, H. (2019). Development and delivery of species distribution models to inform decision-making. *BioScience*, *69*(7), 544–557. https://doi.org/10.1093/biosci/biz045

Song, L., & Estes, L. (2023). *Itsdm* (v0.2.0). *Zenodo*, https://doi.org/10.5281/zenodo.7533022

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Thuiller, W., Georges, D., Gueguen, M., Engler, R., & Breiner, F. (2021). *biomod2: Ensemble platform for species distribution modeling*. https://CRAN.R-project.org/package=biomod2

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD—A platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1). https://doi.org/10.1002/ecm.1486

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1: Supplementary tables, figures, and scripts.
Appendix S2: Evaluation metrics.

How to cite this article: Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, *00*, 1–10. https://doi.org/10.1111/2041-210X.14067