Expert Systems With Applications 206 (2022) 117841

Contents lists available at ScienceDirect

Expert Systems With Applications

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

DeepData: Machine learning in the marine ecosystems

Leonor Oliveira e Silva, Magda Resende, Helena Galhardas, Vasco Manquinho, Inês Lynce *

INESC-ID/IST, Universidade de Lisboa, Rua Alves Redol 9, 1000-019 Lisboa, Portugal

ARTICLE INFO

Keywords:
Machine learning
Species distribution models
Marine ecosystems

ABSTRACT

Based on environmental and species monitoring data, Species Distribution Modelling (SDM) tries to build a model to predict the distribution of a species across a geographic area. These models can then be used to manage the activities in the area in order to prevent negative economic and environmental impacts. In marine ecosystems, SDM can be used to regulate fishing practices or manage protected areas.

This paper presents `DeepData`, a new no-code web-based machine learning platform to facilitate the work of marine biologists with SDM. The `DeepData` tool enables to automate SDM, by automating the creation and validation of the model by marine biologists. Biologists mostly use probabilistic algorithms, such as maximum entropy, generalized linear models and generalized additive models. The `DeepData` tool also allows the use of machine learning algorithms, such as classification and regression trees, random forests and support vector machines. Moreover, besides the usage of machine learning algorithms, other steps in SDM, such as data preparation and model evaluation, are also discussed in the paper. Furthermore, a concrete explanation of the use of the `DeepData` tool is presented, as well as the details of implementation and evaluation.

1. Introduction

The world's oceans face increasing pressure from human influences. Marine ecosystems are utilized by several economic sectors, namely commercial and recreational fishing, tourism and passenger transportation. Species are vulnerable to impacts from all these activities due to competition with fisheries, habitat degradation and disturbance (Koundouri & Giannouli, 2015).

Open ocean and deep sea are under increasing threat from various human activities. The most pressing threats come from overfishing, destructive fishing practices and illegal, unreported and unregulated fishing activities. Other emerging problems include ship-based marine pollution, illegal dumping and noise pollution (IOC-UNESCO and UNEP, 2016a, 2016b).

Given its high levels of biodiversity and wealth of resources, spatial planning is recognized as an essential tool for effective management of all human activities occurring in the deep sea and to ensure a sustainable exploitation of its resources (Guisan et al., 2013). The success of spatial planning and the design of protected areas rely on a good understanding of the spatial distribution patterns of species. Through research and monitoring of species, datasets are created in order to help understanding and managing ecosystems by characterizing the species habitats. With a reliable dataset consisting of locations where species have been observed, a pattern of the suitable conditions of each species can be inferred. As a result, one can try to infer where each species occurs and does not occur without having to sample the whole ocean. Spatial information can then be used to infer the status of the species. Nevertheless, extensive sampling programs for the deep sea are costly and technically challenging, in comparison to shallow inshore waters, where spatial planning is a much easier task (Brandt et al., 2014).

Species distribution modelling (SDM) explores the relations between environment and species, to predict the distribution of species across geographic areas (Elith & Leathwick, 2009b; Guisan et al., 2013). As technology evolves, new methods are proposed for biologists to model species' distributions. Nowadays, the methods most commonly used by marine biologists are based on statistical approaches (e.g. regression), but newer methods based on effective machine learning algorithms are also available.

In the last decade, several platforms that enable non-programmers to build and use complex systems (Iyer et al., 2021; Kölzsch et al., 2022; Schötteler et al., 2021) have been developed. These no-code platforms allow domain experts to manipulate data and develop their applications using complex methods without the need to use a programming language. In order to help biologists with the application of new machine learning algorithms on their data, and as a way to facilitate the comparison with more traditional models, we have developed `DeepData`, a new web-based machine learning tool for marine ecosystems. Nowadays, biologists still have to program the

* Correspondence to: INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal.
E-mail addresses: leonor.oliveira.e.silva@tecnico.ulisboa.pt (L. Oliveira e Silva), magda.resende@tecnico.ulisboa.pt (M. Resende), helena.galhardas@tecnico.ulisboa.pt (H. Galhardas), vasco.manquinho@tecnico.ulisboa.pt (V. Manquinho), ines.lynce@tecnico.ulisboa.pt (I. Lynce).

https://doi.org/10.1016/j.eswa.2022.117841
Received 16 April 2020; Received in revised form 8 June 2022; Accepted 9 June 2022
Available online 17 June 2022

SDM, which can sometimes be hard and prune to errors as it is not their area of expertise. An alternative is to ask a programmer to program the SDM, but there is also a steep learning curve on the part of programmers to learn what they need to develop. Hence, in this paper we propose DeepData, a web-based platform to configure the whole SDM process and visualize the resulting data. On one hand, DeepData facilitates the job of the biologists since they no longer have to spend time programming, and can instead focus all their time on analysing results, which require their expertise. On the other hand, biologists can now easily compare many different SDM models, which otherwise they would have to learn how to program before using them.

The paper is organized as follows. Section 2 defines SDM and describes the different stages of the modelling process, namely (i) data pre-processing; (ii) model selection and training; and (iii) model evaluation. Next, Section 3 describes in detail the implementation of the DeepData tool. Both Sections 2 and 3 follow the same structure for an easy association between the theory and the implementation. Section 4 illustrates the usage and usefulness of the DeepData tool by replicating two different case studies. Finally, the paper concludes in Section 5.

2. Background

SDM assumes that species distributions depend on the physical environment. The concept that the distribution of a species depends on the environment is known as an ecological niche. Therefore, this area of study is also referred to as ecological niche models. According to SDM, species are constrained by their tolerance to environmental factors (Hirzel & Le Lay, 2008).

SDM tries to understand the dependence of each species on the environment conditions. By projecting the environment conditions into geographic space, it is possible to estimate species' geographic distribution, which corresponds to the model. SDM predicts where the species could survive. Furthermore, it is a very useful mechanism to monitor the variations in the habitat suitability of species, impacts of climate change and studies of species delimitation (Porfírio et al., 2014). To characterize the suitable environmental conditions, we use species occurrence data and environmental data. By interpolating species occurrences and environmental datasets, it is possible to find a pattern that describes the environmental conditions. The model usefulness and robustness are influenced by the selection of variables and modelling methods and how the relation between environmental and geographic factors is handled (Elith & Leathwick, 2009a).

The SDM creation process is composed of three main steps: (i) data pre-processing; (ii) model selection and training; and (iii) model evaluation. The process is represented in Fig. 1. The data pre-processing phase is represented in red, model selection and training are represented in blue and model evaluation is represented in green.

By examining the figure, we observe that SDM requires that each step is performed multiple times as evaluation is performed and feedback is obtained, thus leading to a better fit of the model. Note that the detailed steps to build a SDM are not fixed. Only the main steps that serve as guidelines are defined.

2.1. Data pre-processing

During SDM, we relate occurrence data with environmental data that is thought to determine the species distribution. Therefore, the process assumes that the occurrence's data covers the species' full ecological range. One of the challenges of this process is to have enough occurrence records, as well as accurate and relevant environmental variables at a sufficiently high spatial resolution. The user has to take some precautions while choosing the data, in order to guarantee the reliability of the data.

Regarding occurrence data, the coordinates of the location data need to be accurate so that the species/environment association is reliable. Even taking into account this precaution, occurrence data might be biased towards the accessibility of sampling locations as data may be lacking for remote areas.

There are two types of occurrence data: presence-only and presence/absence. Presence-only refers solely to the location of where the species are present. Presence/absence refers to when we have the location of both where the species are present and where they are absent. When dealing with absence, we have to be careful because it can mean that habitat is unsuitable or it is suitable but unoccupied (maybe because it is inaccessible). Absence data is also tricky to get because the fact that a species is not detected in a location at a given time does not mean it does not exist there.

Environmental data need to be in grid type format, where each environmental variable is divided into grid cells representing its value for a location at some resolution. The selection of the resolution has to take into account spatial autocorrelation[1]. Spatial autocorrelation states that the closer two locations are, the more similar are their measures of species occurrences (Carsten F. Dormann, 2007; Crase, Liedloff, & Wintle, 2012). This similarity is due to biotic processes, such as reproduction, predator–prey interactions, food availability, etc. This similarity phenomenon leads to dependence among samples decaying with distance, which violates the assumption of data independence. It also leads to the underestimation of variance and the overestimation of the significance of effects.

The DeepData tool facilitates the work of the user by enabling the calculation of pseudo-absences, which are useful if no absences data is known. Pseudo-absences are inferred based on presence data. DeepData allows the user to measure Moran's I, which evaluates spatial autocorrelation. DeepData also measures the collinearity between variables with the variance inflation factor. Collinearity is the most used measure in the context of SDMs. It refers to the existence of correlated environmental variables, which can lead to biased models due to inflated variances. Small changes in the data set can strongly affect results and so the SDM tends to be unstable (high variance) and the relative importance of the variables is difficult to assess (Dormann et al., 2013).

The success of the prevision depends mostly on the quality of the data used, both of species and environment because it cannot be biased and is the base of the learning process. Therefore, pre-processing is the part that takes the longest, and is performed various times along the whole process.

2.2. Model selection

Prediction models need to balance specific fit to the training data against the generality that enables a reliable prediction for new cases.

For the modelling phase, the DeepData tool enables cross-validation to be performed. DeepData allows the user to select one of the following methods:

- Holdout, which separates the dataset into train set and test set according to a given fraction, being the train set larger than the test set. Training is performed as many times as there are test sets.
- Leave one out, which separates the data into three partitions and at each repetition uses two for training and one for testing.
- K-fold, which separates the data in k folds and trains the model over the k number of combinations.
- Years separation, which separates the data according to the years selected for train and test, with the constraint that each year can only be either train or test.

[1] Spatial autocorrelation can be assessed through Moran's I measure (Moran, 1950).

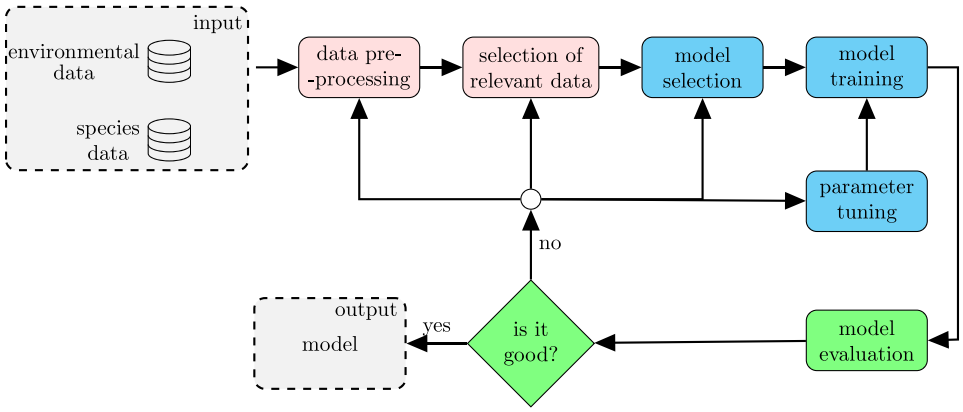


Fig. 1. Representation of SDM.

For all the methods, the considered statistical algorithms are: Generalized Linear Models (`GLMs`) (Dobson, 2010), Generalized Additive Models (`GAMs`) (Hastie & Tibshirani, 2005) and Maximum Entropy (`MaxEnt`) (Phillips, Dudík, & Schapire, 2004). Moreover, the considered machine learning algorithms: are Random Forests (`RFs`) (Breiman, 2001), Support Vector Machines (`SVMs`) (Cortes & Vapnik, 1995) and Artificial Neural Networks (`ANNs`) (Palm, 1986).

Depending on the data, some models may be more adequate than others. While `MaxEnt` is a presence-only model, the remaining ones are presence–absence models. Both `GLMs` and `GAMs` are statistical models that fit an equation to data. `RFs` combine various decision trees, which recursively divide the variable space into two smaller partitions. `SVMs` visually create a separation of data. Finally, `ANNs` are inspired by the capacity of the human brain to learn.

If more than one model is selected, then an ensemble model is computed additionally to the models selected. Ensemble models (Zhou, 2012) use multiple learning models to obtain a better predictive performance than the performance of a single model. A single model can have biases and inaccuracies that affect reliability. By combining the decisions of different models, these effects can be reduced, thus improving the overall performance. This is due to the fact that correct answers are reinforced while incorrect ones tend to be blended. This ensemble can be performed by:

- Voting,
- Averaging,
- Weighted.

In voting, each model outputs a prediction for each data point. Each of these predictions is considered a vote. The final prediction, meaning the prediction of the ensemble, corresponds to the majority. The averaging method is similar to maximum voting, but instead of the final prediction being the majority, it is the average of all the single predictions. In weighted, instead of returning a simple average of the predictions, it assigns a weight to each prediction. This weight defines the importance of each model, which can be accessed through some evaluation metric, such as accuracy, kappa statistic, sensitivity, specificity or proportion correct.

2.3. Model evaluation

Assessing the model performance (Allouche, Tsoar, & Kadmon, 2006; Guisan & Zimmermann, 2000; Pearce & Ferrier, 2000) helps determine its suitability and the aspects that need improvement. Without a relevant evaluation, the model has no value. Moreover, it also enables to compare different models.

We can use two types of evaluations: (i) threshold dependent metrics; (ii) threshold independent metrics; and (iii) visual metrics.

Regarding threshold dependent metrics, performance can be assessed by selecting a threshold and constructing a confusion matrix. A threshold must be defined to classify an instance either as presence or absence. From the confusion matrix, various measures of performance can be derived such as accuracy, sensitivity, specificity and kappa statistic (Liu, White, & Newell, 2011).

The accuracy measures the proportion of correctly predicted instances. The problem with accuracy is that it is prevalence sensitive. Other measures, such as sensitivity, which measures the proportion of observed presences that are predicted as such, and specificity, which measures the proportion of observed absences that are predicted as such, are independent of prevalence (Manel, Williams, & Ormerod, 2001). The Kappa statistic assesses the extent to which models predict the occurrence at a rate higher than expected by chance.

Regarding threshold independent metrics, the model can be evaluated by the Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). MAE measures the average magnitude of the error, without considering their distance. MSE and RMSE give more weight to a large error, because of the square. The fact that RMSE uses the root makes it easier to interpret its value.

Regarding visual metrics, we can use the residual vs fitted values plot. A residual is the difference between the observed and the predicted. If the residual is not random, then it means that something is missing in the model. This plot tests the assumptions of whether the relationship between the variables is linear and whether there is an equal variance along the regression line. Partial residuals represent the residual of a variable in regards to the occurrences after subtracting the contribution of the other variables.

Another visual tool is the Quantile–Quantile plots (QQplots), which allows us to assess if the data follows some theoretical distribution. To visually assess spatial correlation, scale location plots are used. These plots show if residuals are spread equally along with the ranges of predictors. Residual leverage plots are used to assess influential data points, i.e points whose inclusion or exclusion produce different results in the model. Finally, partial dependence plots allow the visualization of the relationship between the occurrences and each environmental variable, while accounting for the effect of the other variables.

The Relative Operating Characteristic (ROC) curve (Jiménez-Valverde, 2014) is a graphical measure that describes the trade-off between sensitivity and false positive as the decision threshold varies. False positive measures the proportion of observed absences that are predicted as present. By knowing the number of observed presences, if we have the sensitivity measure we know the number of correctly predicted presences and we infer the number of absences that were incorrectly predicted as presences. Similarly, by knowing the number of observed absences, if we have the false positive measure we know the number of incorrectly predicted presences and we are able to deduce the number of correctly predicted absences. Therefore, the ROC Curve describes the whole model.

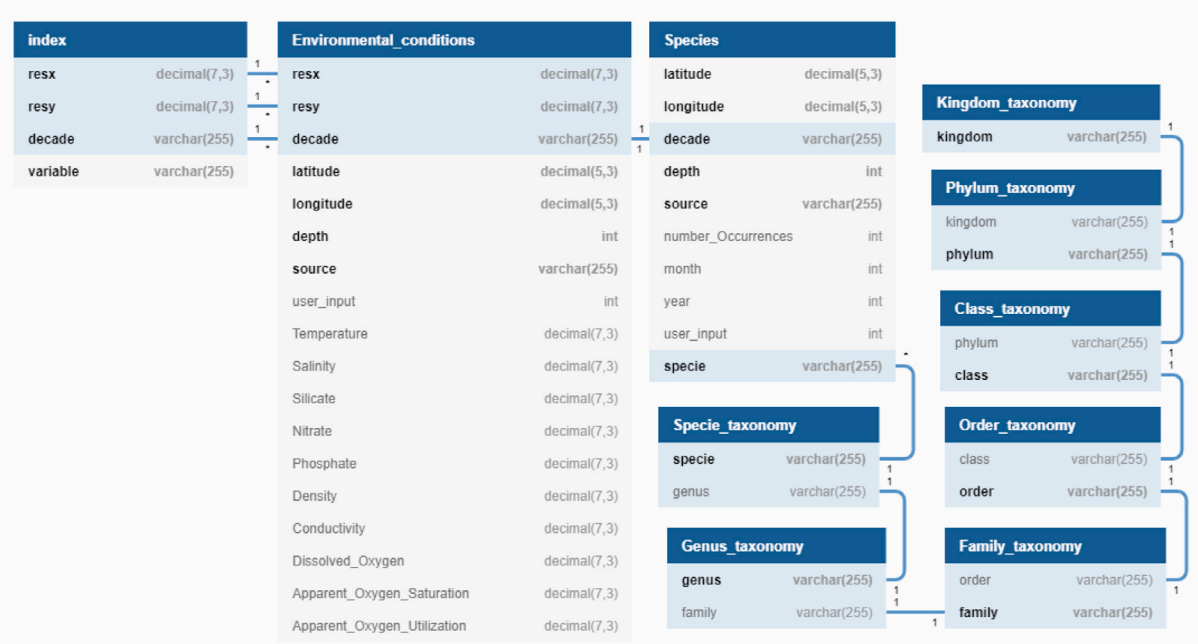


Fig. 2. Database architecture scheme.

The Area Under the ROC Curve (AUC) is used as a global metric predicting the overall discriminatory ability of the model. AUC gives the probability that a randomly chosen presence site will be ranked above a randomly chosen absence site. When no absence is available, such as with MaxEnt, then the AUC gives the probability that a randomly chosen presence site will be ranked above a randomly chosen background site.

3. The DeepData tool

DeepData started as a project to help researchers of IMAR[2] to study the Azores deep sea in an automatic way. The motivation for building the tool was to overcome the need for researchers to code a new model every time new data is available. For some species, researchers know the most suitable model to apply and with DeepData they can simply run the model through the tool. Even if the optimal model is not known, the tool can be used to test different models. DeepData was later extended to cover all marine ecosystems.

3.1. Database architecture

DeepData includes data from World Ocean Atlas 2013, European Marine Observation and Data Network, Ocean Biogeographic Information System and World Register of Marine Species datasets, concerning the geographic space of the Azores EEZ. The environmental variables in these data are commonly used to model species distribution models. As a result, the user can make experiments with DeepData without having to insert data. Concerning the World Register of Marine Species dataset, it is not used to model the species distribution but rather to give information about the taxonomy of the species that is being modelled.

DeepData's database is composed of the following tables, as shown in Fig. 2:

- `Index`, which stores the variable names and their `resolution on x`, `resolution on y` and `decade` available.
- `Environmental_conditions`, which stores the environmental variables (temperature, salinity, silicate, nitrate, phosphate, density, conductivity, dissolved oxygen, apparent oxygen saturation and apparent oxygen utilization) associated to their `latitude`, `longitude`, `depth`, `source`, `decade`, `resolution on x` and `resolution on y`.
- `Species`, which stores data about the species' occurrences. It associates the `species name` from table `Kingdom_taxonomy` with its `latitude`, `longitude`, `depth`, `source`, `decade`, `year` and `month`.
- `Kingdom_taxonomy`, which stores the species' `kingdom`.
- `*_taxonomy`, which stores the association between the species `kingdom-phylum-class-order-family-genus-species`.

In the case that the researcher wants to use private data, DeepData allows the insertion of both species and environmental data, as shown in Fig. 3. Species data can be uploaded using a `csv` file, with the following structure: `species' name, latitude, longitude, month collected, year collected, depth, data source name`. Environmental data can be uploaded in a `csv` or `ascii` format. `Ascii` files are composed of a header that defines the properties of the file, such as:

- Number of cell columns, represented by `ncols`.
- Number of cell rows, represented by `nrows`.
- Longitude coordinate of the origin, by centre or lower left corner of the cell. Represented by `xllcenter` or `xllcorner`, respectively.
- Latitude coordinate of the origin, by centre or lower left corner of the cell. Represented by `yllcenter` or `yllcorner`, respectively.
- Cell size, which might have two values, if the cell sizes for the longitude and latitude are different. Represented by `cellsize`.
- The values that mean no data, represented by `nodata_value`.

The header is followed by cell value information, separated by a space character. The `csv` files to upload the data require the following structure: `latitude, longitude, depth, decade, environmental variable value, data source name`, where decade has the format year1_year2.

[2] https://imar.org.pt/en/about-us/.

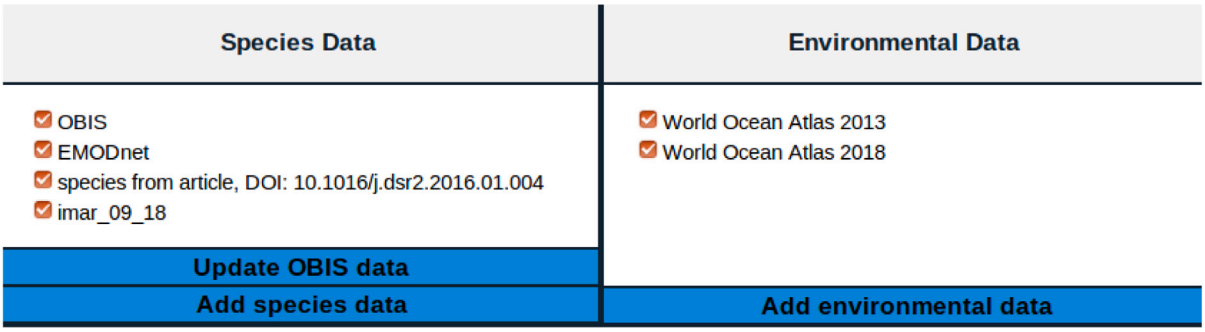


Fig. 3. DeepData's data insertion interface.

3.2. Tool implementation

For the implementation of the species distribution models, the software used was R.[3] Although `python` is also commonly used for machine learning, most of the published work on SDMs uses the R software. Ecologists tend to use R while computer engineers tend to use `python`, because R has been established for a long time and includes a broader range of methods employed in the ecological analysis as well as numerous routines for data exploration (Lai, Lortie, Muenchen, Yang, & Ma, 2019). `Python` has the advantage of being better for deployment, and therefore it is used to implement other parts of the application, including fetching the data from the database needed for computing the SDM.

Fig. 4 presents the `DeepData`'s user interface, which allows the user to select the parameters she wants to apply.

3.2.1. Data pre-processing

`DeepData` enables the insertion of species data through a csv file. Each entry in the csv file is then associated with a decade so that the data stored in the table `Species` is associated with the data stored in the table `Environmental_conditions`. Moreover, every new species inserted in table `Species` is also inserted in table `Species_taxonomy` so that the data in both tables is related. To insert environmental data, the user can upload `csv` files, but also `ascii` files. For both files, the user has to specify whether the variable to be inserted already exists.

The species can be selected through the taxonomy hierarchy, or by directly selecting the species name.

Data pre-processing can be manually done by the user or can be automatic, meaning that `DeepData` automatically does it. Regarding manual data pre-processing the user can generate pseudo-absences, select oceanic and terrain variables and calculate Moran's I correlation coefficient. Regarding automatic pre-processing, `DeepData` automatically unifies different variables and measures the collinearity between variables.

The user can choose either to generate pseudo-absences or not. To generate pseudo-absences, `DeepData` starts by classifying the geographic space as suitable or unsuitable according to the environmental conditions of the presence localities. Then, the pseudo-absences are randomly generated in the unsuitable environment, having the same proportion of presences, meaning a prevalence of 0.5 and a minimum distance to the presences of 30 km.

For the environmental variables, the user can select oceanic variables, which can have different resolutions, and terrain variables. `DeepData` also enables the selection of various oceanic and terrain variables, and the definition of the oceanic zone. The oceanic zone can be:

- Ocean surface, meaning that only a depth of 0 to 200 meters is considered.
- Ocean floor, meaning that only a depth of 4000 to 6000 meters is considered.
- Average depth of the species occurrence, meaning depth values are prioritized by the number of occurrences of the species.

The range of the ocean floor values is much larger than the range of the ocean surface values since spatial variation in environmental variables decreases with depth (Costello, Basher, Sayre, Breyer, & Wright, 2018). When the `average depth of the species occurrence` option is selected, `DeepData` computes a ranking of depth by the frequency of the species. For example, if a species appears 40 times at depth 200, 30 times at depth 350 and 10 times at depth 20, we first load all variable values at depth 200. For the cells where data at this depth is not available, we load the values for a depth of 350 and finally for the remaining cells the depth of 20. An average could have been computed instead, but by analysing the species' occurrence data, we conclude that species might occur at any depth, due to the bathymetry variation.

There is also the possibility to calculate the Morans'I correlation coefficient. Given the environmental data, Morans'I evaluates whether the pattern expressed is clustered, dispersed or random. A clustered spatial pattern means most of the values are concentrated in nearby locations or adjacent together. A random spatial pattern means the distribution of the values is homogeneous or independent. A dispersed spatial pattern means that values similar to each other are located far from each other in a uniform way.

Now `DeepData` needs to unify the different variables. When the variables have different resolutions, `DeepData` uses the highest resolution as the standard resolution, i.e., the finer resolution. As shown in Fig. 5, the aggregation of the original variable by a factor 2, meaning each cell covers 1/4 of the area, is equal to the mean of the four original cell values. In Fig. 5 are represented on the left, in blue (grey), four cells, whose aggregation will correspond to their mean, are represented, with blue (grey), in the middle figure. The Fig. 5 also shows the disaggregation of the aggregated variable by a factor of 2, as an attempt to reproduce the original variable. So the table on the right represents, in blue (grey), four cells, corresponding to the disaggregation of the blue (grey) cell in the middle figure. As shown, the disaggregation corresponds to the expansion of the cell to the desired extent. Now, each cell covers 4 times the area. Therefore, `DeepData` uses the lower resolution, because aggregating leads to information loss while disaggregating has no effect on the amount of information.

`DeepData` also measures the collinearity between variables. Collinearity refers to the existence of correlated environmental variables, which can lead to biased models due to inflated variances. Small changes in the data set can strongly affect results so the SDM tends to be unstable (high variance) and the relative importance of the variables is difficult to assess (Dormann et al., 2013). Only checking the collinearity between pairs of variables can be limiting, so the Variance Inflation Factor (VIF) quantifies the extent of correlation between one variable

[3] https://www.r-project.org/.

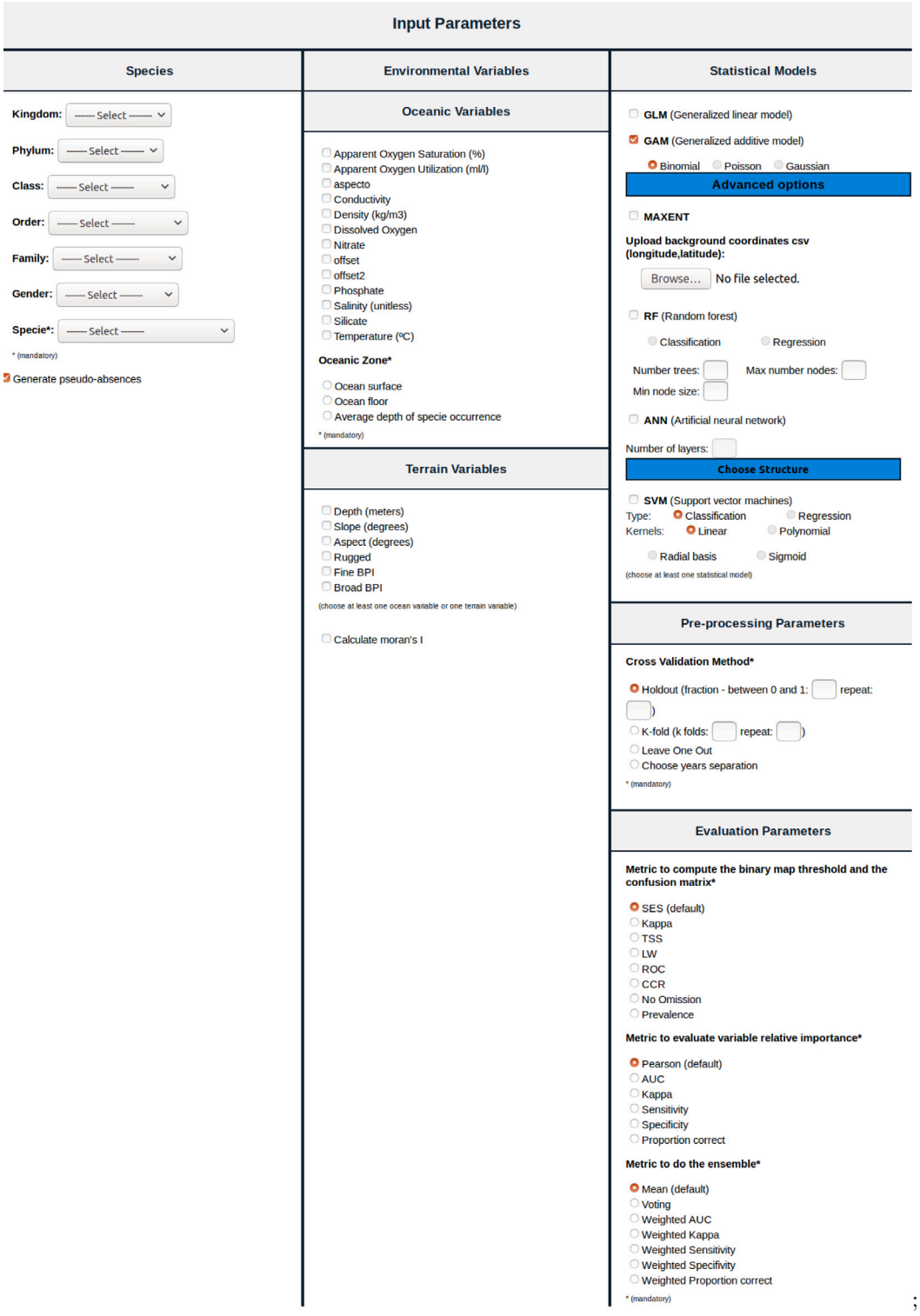


Fig. 4. DeepData's input interface.

and the other remaining variables. For variance inflation factors larger than 3, which means that the standard error is 1.7 times larger than if the variables were not correlated, the modelling process stops and a pop-up appears asking the user whether he/she wants to remove any variable. If the user chooses to maintain all the variables then the collinearity is not verified again. If the user chooses to remove some variables, then the collinearity is verified for the remaining ones.

3.2.2. *Model selection and training*

For the modelling phase, DeepData supports cross-validation. DeepData allows the user to select one of the following methods:

- Holdout,
- Leave one out,
- K-fold,
- Years separation.

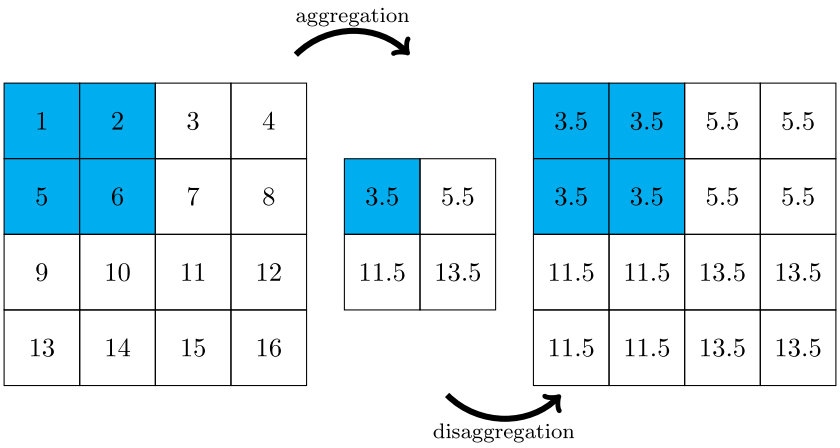


Fig. 5. Example of aggregation and disaggregation.

Table 1
R packages used for each implemented model.

Model	R package
`Generalized addictive model`	`mgcv`
`Generalized linear model`	`SSDM`
`MaxEnt`	`dismo`
`Random Forest`	`randomForest`
`Artificial neural network`	`neuralnet`
`Support vector machine`	`e1071`

Table 2
Relation between the family distribution and the link function.

Family distribution	Link function
Binomial	logit probit cauchit log cloglog
Possion	log identity sqrt
Gaussian	log identity inverse

Moreover, `DeepData` enables the specification of the model parameters. Each model is implemented with a specific R package, from the list shown in Table 1.

In order to compute Generalized Addictive Models, `DeepData` enables the specification of the family of the distribution, which can be:

- Binomial, in the case of presence–absence data, which is used as default.
- Poisson, in the case of count data.
- Gaussian, in the case of count data with a normal distribution.

It also supports the specification of the link function, in accordance with the family, as shown in Table 2. Smoothness of fit of each variable can also be controlled, differing on the basis used to represent the smooth function. Possible splines are: (i)thin plate spline, which is used as default, (ii) duchon spline; (iii) cubic spline; (iv) spline on the sphere; (v) p-splines; (vi)random effects; and (viii) no smoothing.

For computing Maximum Entropy, `DeepData` enables uploading a background file, which must be composed of the latitude and longitude. If no file is given, `DeepData` randomly selects 10000 points of the coordinate space to be used as background.

Table 3
Threshold metric and definition.

Metric	Definition
SES	Maximizes sensitivity equal to specificity
Kappa	Threshold value or range in values with the maximum Kappa statistic
TSS	Maximizes sensitivity plus specificity
LW	Minimizes prediction probability for the occurrence (presence) records
ROC	ROC curve is closest to point (0,1)
CCR	Maximizes number of presence and absence records correctly identified
No omission	No false positives (predicting absences incorrectly)
Prevalence	Modelled prevalence closest to the observed prevalence

For computing Random Forests, `DeepData` supports tuning the following parameters:

- Number of trees to grow. This should not be set to a number too small, to ensure that every input row gets predicted at least a few times. Default is set to 500.
- Minimum size of the terminal nodes. Setting this parameter to a large number causes smaller trees to be grown (and thus take less time). The default values are 1 for classification and 5 for regression.
- Maximum number of terminal nodes that the trees can have. If not given, trees are grown as much as possible (subject to limits by node size).

For computing Neural Networks, the structure can be defined by first indicating the number of layers, which corresponds to the sum of the input layer, hidden layers and output layers. Afterwards, the number of perceptrons for each layer is defined. Note that Neural Networks do not have any default structure.

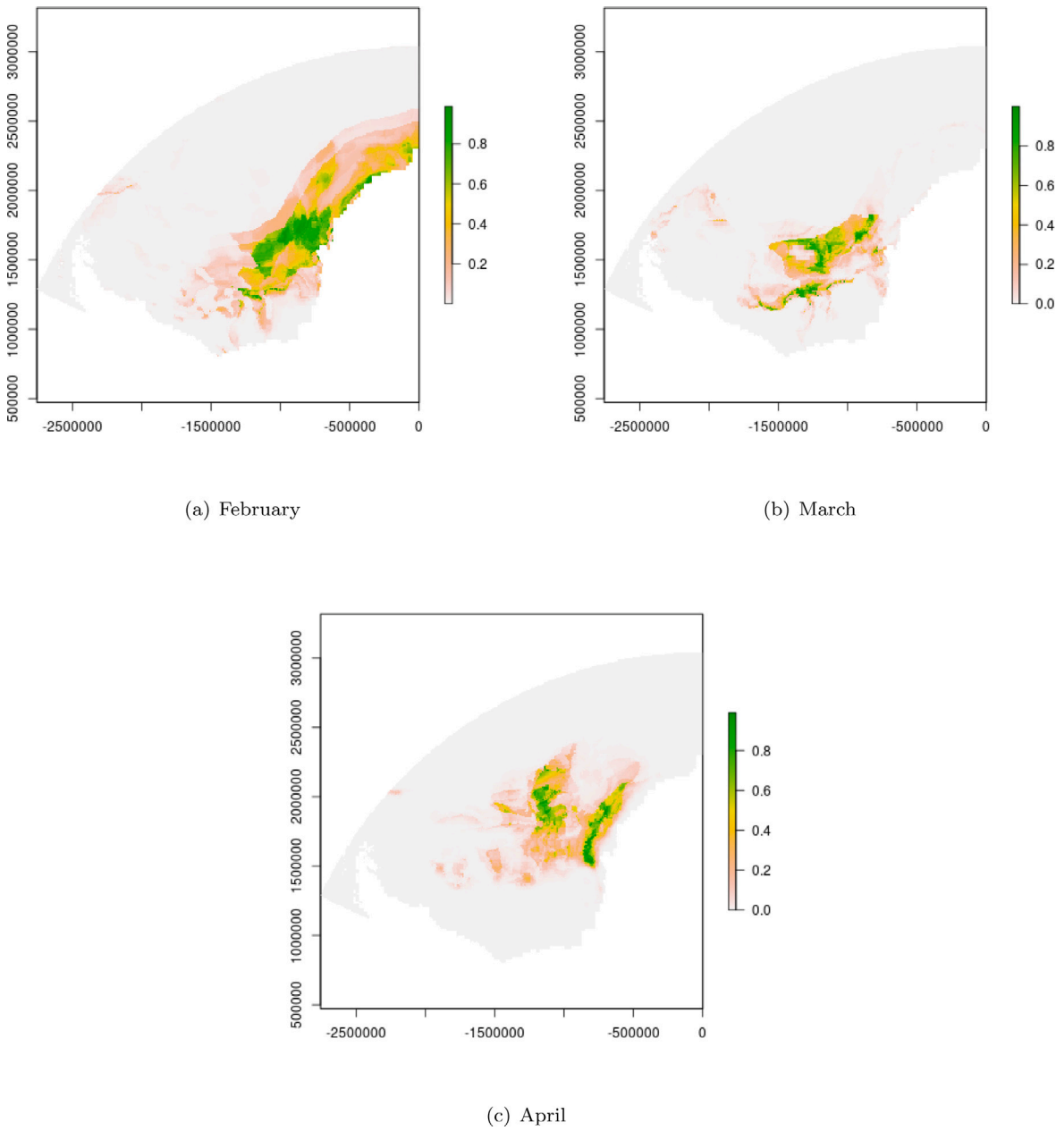
For computing Support Vector Machines, the kernel can be defined as: (i)linear, which is the default value; (ii) polynomial; (iii) radial basis; or (iv)sigmoid.

If more than one model is selected, then an ensemble model is computed additionally to the models selected. This ensemble can be performed by: (i) averaging; (ii)voting; or (iii) weighted.

3.2.3. Model evaluation

The user has to select the metric to compute the binary map threshold and the confusion matrix. `DeepData` enables the threshold to be defined by various metrics, which are shown on Table 3.

While the first three metrics (SES, Kappa and TSS) can be applied to all models, the last metrics (No Omission and Prevalence) can only



(a) February (b) March

(c) April

Fig. 6. Probability of presence of crabeater seals for each month.

be applied to `MaxEnt`. The remaining metrics (LW, ROC and CRR) can be applied to all models except `MaxEnt`, i.e. `GAM`, `GLM`, `RF`, `ANN` and `SVM`.

For all models, `DeepData` returns the threshold, accuracy, omission rate, sensitivity, specificity, proportion of correctly predicted occurrences and kappa statistic of the best model. To access the overall variation of each model of the cross validation, the mean accuracy, mean threshold and corresponding standard deviations are presented. It also returns the calculated VIF (variance inflation factor) of each variable and the number of presences used.

`DeepData` returns a folder, with model evaluation plots, specific to each model. In the folder, there is always a file with the plot of the predicted occurrence values over the environmental space.

For the `GAM` model, `DeepData` also returns a residuals analysis, the component smooth functions, standard error estimates and Akaike's Information Criterion.

For the `RF` model, `DeepData` also returns each variable importance and the effect of a variable on the predicted occurrence.

For the `MaxEnt` model, `DeepData` also returns the ROC curve and information about the statistical significance of the prediction and analysis of variable contribution is provided.

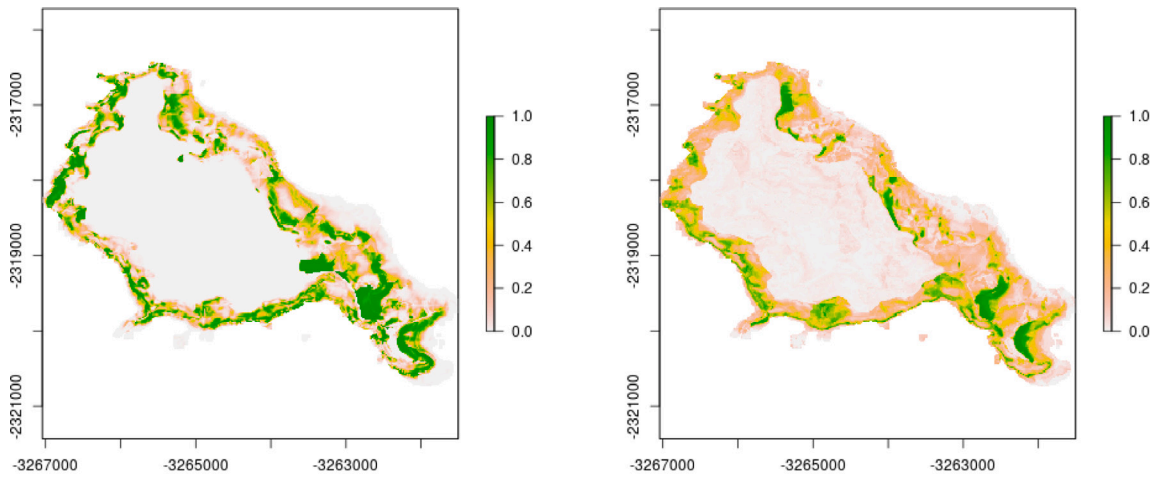
4. Case studies

In order to demonstrate the usefulness of the `DeepData` tool, we selected two different research works with publicly available data and used `DeepData` to replicate their results. The first case study is related to the inference of the habitat conditions of the crabeater seals in the Weddell Sea. In the second case study, the habitat conditions of the Trindade Petrel of the Brazilian coast is studied.

4.1. First case study: Crabeater seals

4.1.1. Problem description

The crabeater seals is one of the most abundant Antarctic species that inhabits the hardly accessible Antarctic pack ice zone, in particular



(a) GAM (b) RF

Fig. 7. Spatial distribution of Trindade Petrel for each model.

the Weddell Sea. Due to the lack of accessibility, abundance estimates of the crabeater seals are hard to obtain.

This case study (Nachtsheim, Jerosch, Hagen, Plötz, & Bornemann, 2017) was published in 2017 and is based on data from fifteen crabeater seals of both sexes and different age classes that were equipped with satellite-linked dive recorders (SDRs). During the period of data collection, the sea ice cover of the Weddell Sea was exceptionally low compared to previous years. The impact of this adversity on crabeater seals is not clear, but it could result in a reduction of their habitat. This is becoming more important since the sea ice cover in the Southern Ocean is predicted to decrease due to climate changes.

The researchers also collected data on several environmental variables. In particular, 13 environmental variables were used to analyse the habitat preferences of crabeater seals: sea ice concentration, sea ice thickness, sea ice freezing rate, water surface and bottom temperature, surface and bottom salinity, surface and bottom zonal current velocity, surface and bottom meridional current velocity, slope, and distance to shelf break. All data is available at the `PANGAEA` repository [dataset] (Nachtsheim, Jerosch, Hagen, Plötz, & Bornemann, 2015).

After the data collection, the researchers used `MaxEnt` to model the influence of certain environmental variables on the distribution of crabeater seals in the Weddell Sea. The main goal was to identify the suitable habitat conditions for the crabeater seals. However, before building the model, the seal location data was subsampled to decrease potential biases.

4.1.2. `DeepData` output

In order to replicate the crabeater seal case study, both species occurrence and environmental data of each month were loaded into the `DeepData` database. Next, the `DeepData` tool was configured using the same options as described in the case study (Nachtsheim et al., 2017). The used configurations were as follows:

- Not generate pseudo-absences, since `MaxEnt` only uses presences.
- Use average depth of species occurrence.
- Not calculate Moran's I coefficient.
- Use `MaxEnt`, with the background file loaded.
- Use holdout with a fraction of 80% and a repeat of 20.

`MaxEnt` was used with all environmental variables loaded to verify the contribution of each environmental variable to the model by a measure called permutation importance. As a result, we can identify the environmental variables that were the most relevant concerning the seal distribution. As shown in Table 4, from the 13 environmental variables, slope, bottom zonal current velocity and bottom meridional current velocity did not contribute more than 5% to any monthly model. Therefore, these variables were omitted from the final model. Moreover, it can also be seen that distance to shelf break and sea ice concentration are important environmental variables for determining crabeater seal distribution throughout all three months. Additional environmental variables with moderate overall importance to the models were velocity meridional surface (February) and velocity zonal surface (March), as well as water temperature surface (April).

Table 4
Variable permutation importance in February, March and April.

Variable	February	March	April
Distance to shelf break	41.6	20.1	42.8
Sea ice freezing rate	0.8	0.3	4.2
Sea ice thickness	0.9	0.3	4.6
Sea ice concentration	11.7	35.3	7.7
Salinity bottom	1	6.2	1
Salinity surface	1.4	7.7	3.2
Water temperature bottom	8	3.3	1
Water temperature surface	8.3	10	17.4
Velocity meridional bottom	1.3	0.4	0.5
Velocity meridional surface	22.8	3	11.6
Velocity zonal bottom	0.1	0.7	1.4
Velocity zonal surface	1.4	11.8	4.4
Slope	0.6	1.1	0.2

Table 5
Monthly model AUC and standard deviation.

Month	AUC
February	0.93 ± 0.005
March	0.96 ± 0.006
April	0.94 ± 0.005

Table 5 shows the Area Under the ROC Curve (AUC) values for each month, as well as its standard deviation. As can be observed, all AUC values are high, showing that the predictions are far from random. Moreover, standard deviation values are low, meaning there is a high degree of uniformity between the repetitions.

In order to show that `DeepData` replicates the work of researchers in an automated way, we can compare the final model predictions, shown in Fig. 6, with the ones obtained in the original case study (Nachtsheim et al., 2017). The `x` and `y` axis in Fig. 6 represent latitude and

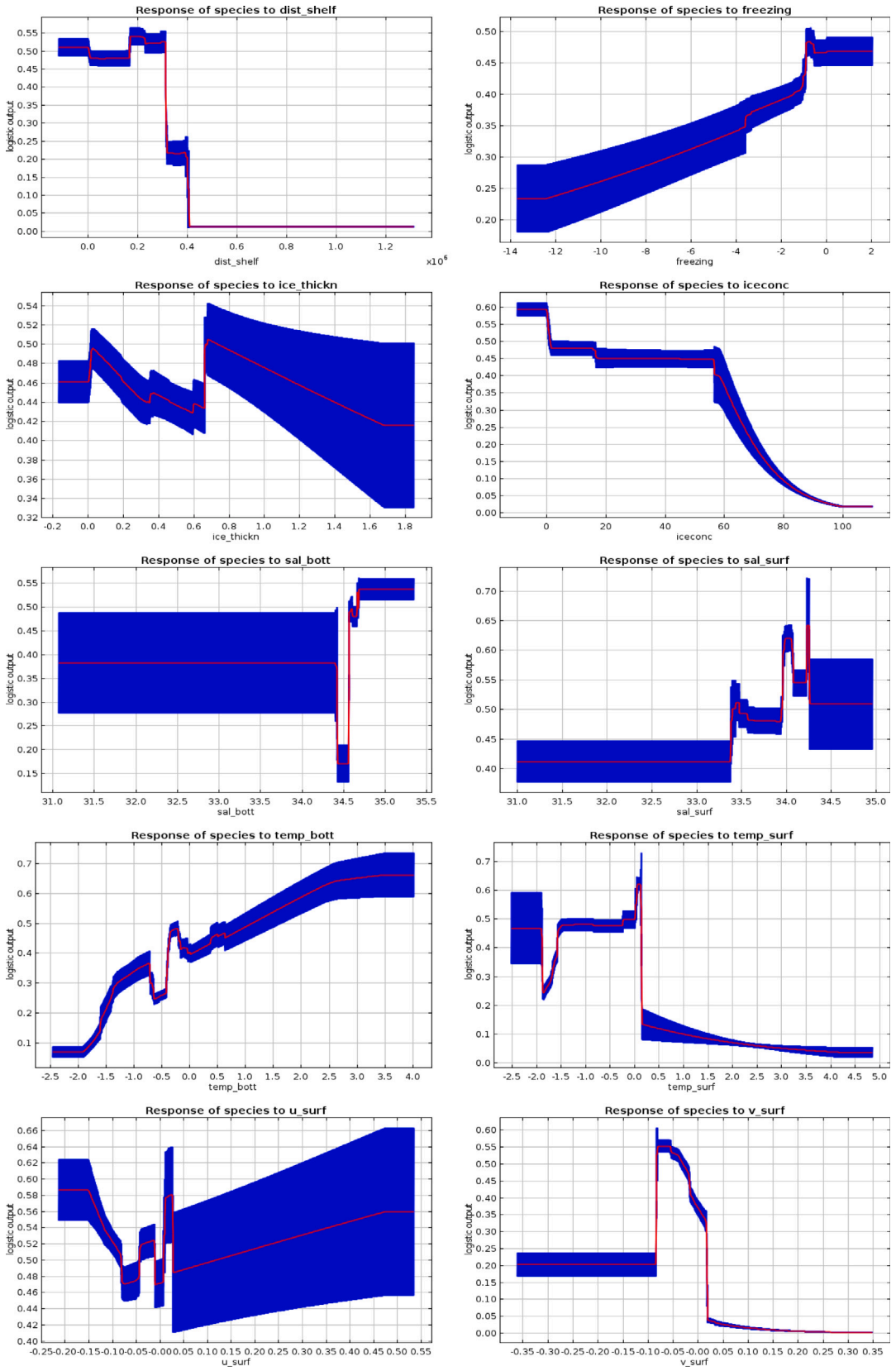


Fig. 8. Response curve for February.

longitude in the projected coordinate system. Furthermore, `DeepData` also generates the response curves, which are shown in detail in Appendix A (Figs. 8–10), and which display how the probability of presence changes with the value range of each environmental variable. These response curves are very similar to the ones in the original paper showing the suitability of `DeepData`.

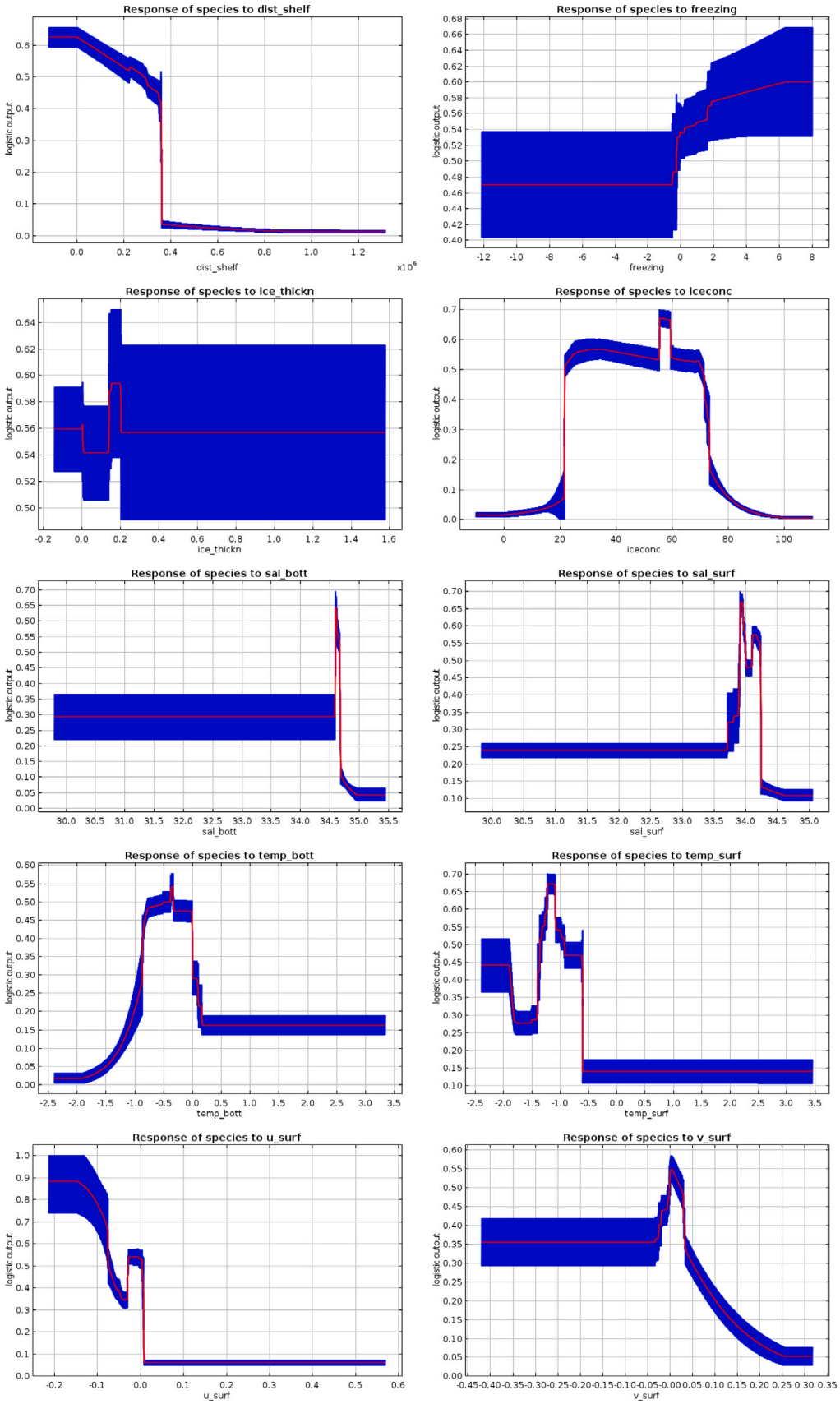


Fig. 9. Response curve for March.

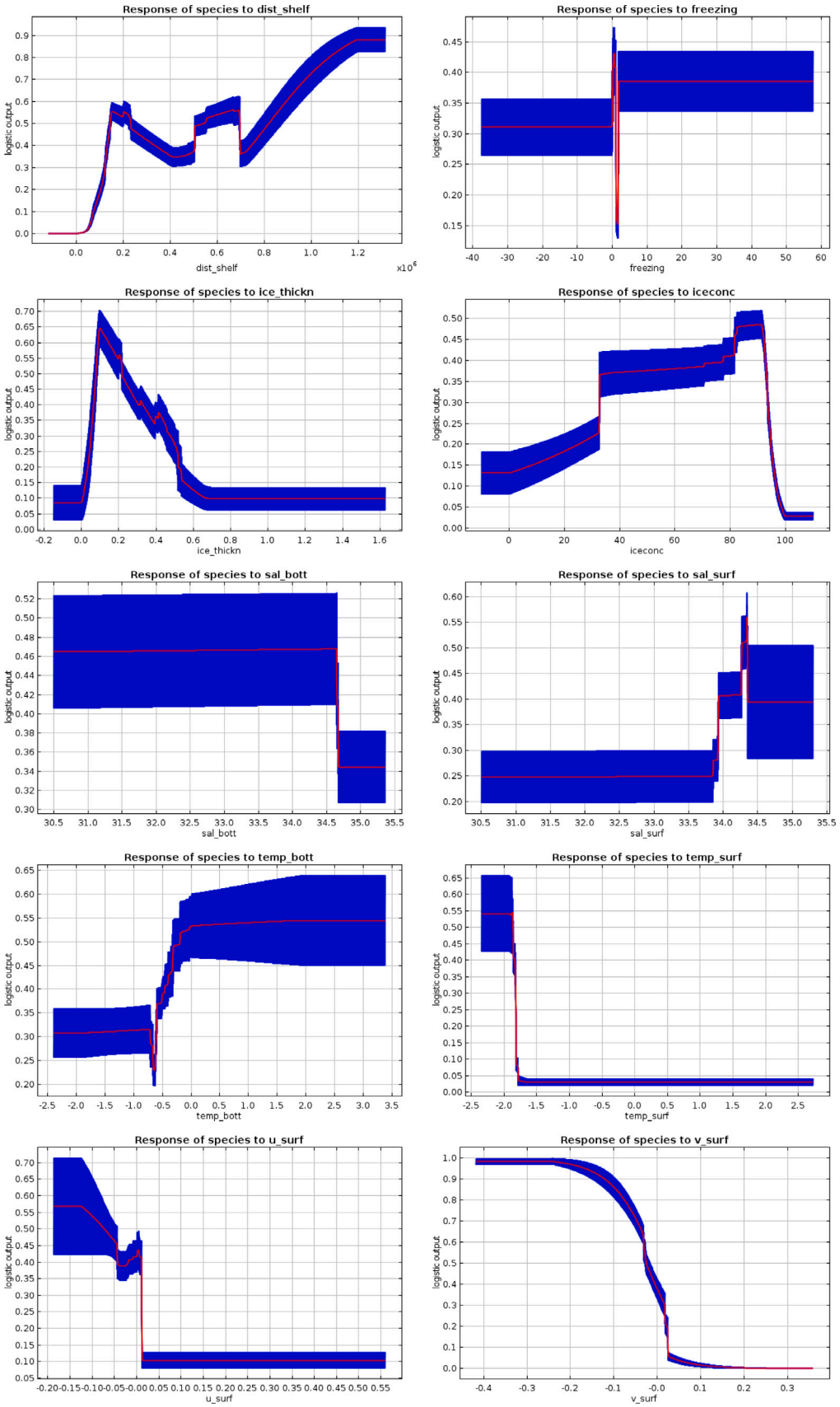


Fig. 10. Response curve for April.

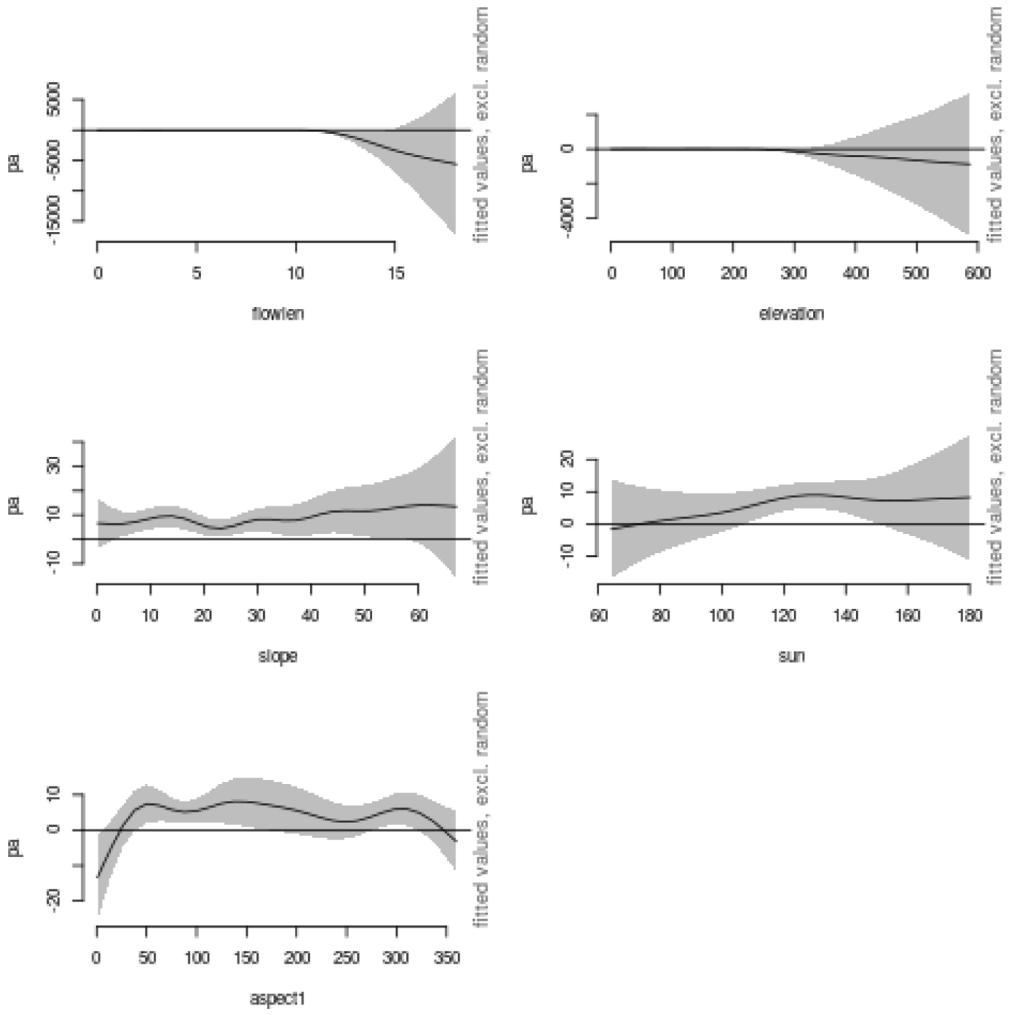


Fig. 11. Partial dependence plot for the generalized additive model.

4.2. Second case study: Trindade petrel

4.2.1. Problem description

The second case study (Krüger, 2018) tries to identify suitable habitat conditions to the Trindade Petrel, which is an endangered gadfly petrel breeding of a Brazil oceanic island.

SDMs applied to seabird studies have been mostly used to predict at-sea distributions, while colony prediction had not been fully explored. However, other methods have been helpful in estimating occurrences in inaccessible seabird species nesting. This particular study (Krüger, 2018) focuses on the use of SDMs to model nest distributions as a way to identify the habitat suitability for seabird colonies nesting in non reachable areas. The paper uses data from a previous study (Krüger, Paiva, Petry, Montone, & Ramos, 2018), while testing a different approach. All data is available at the `PANGAEA` repository [dataset] (Krüger, 2018). Contrary to previous works, an ensemble method is used in this study, since the researchers took into account that different models can produce different distributions.

This research work on the Trindade Petrel was based on 411 nests that were identified over several years. Regarding environmental variables, the following were used: elevation, slope, flow length, aspect and insulation.

In order to create the ensemble model, we tested each individual model. Next, the 3 models that best fit the data were selected to be used in the ensemble model. The tested individual models were the following: `GAM`, `GLM`, `Multiple adaptive regression splines`, `RF`, `Generalized boosted model`, `ANN`, `MaxEnt Phillips` and `MaxEnt Tsuruoka`. The difference between the two variants of `MaxEnt` is the package that implements each of them. While `MaxEnt Tsuruoka` only uses an `R` package, `MaxEnt Phillips` uses a java software which is called within an `R` package. Therefore the modelling options and default parameters are different (*BIOMOD Modeling Options*, 2016).

4.2.2. DeepData output

In order to recreate the study results, both the species occurrence and environmental data were loaded into the `DeepData` tool. Next, the following configurations were used:

- Generate pseudo-absences, since only `MaxEnt` uses presences.
- Use average depth of species occurrence.
- Not calculate Moran's I coefficient.
- Use `GLM`.
- Use `GAM`, with binomial family.
- Use `MaxEnt`, with default values.
- Use `RF`, with classification and min node size of 5.
- Use `ANN`, with 1 layer with 8 nodes.
- Use holdout with a fraction of 80% and a repeat of 3.

This process is repeated 20 times to ensure a generation of different pseudo-absences and different combinations of training and testing datasets. This way, we prevent biases and have a clear idea of which are the best models.

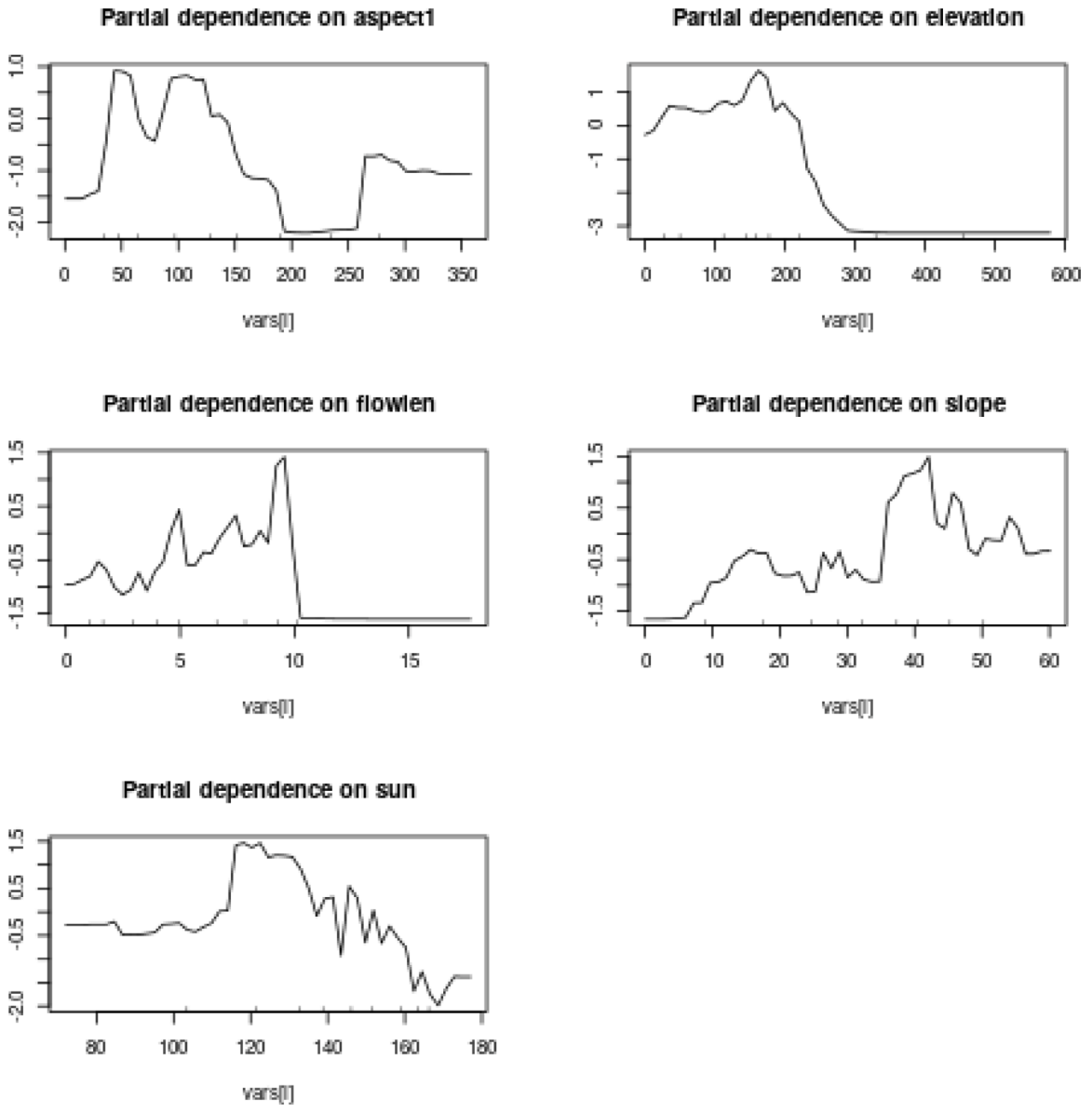


Fig. 12. Partial dependence plot for the random forest model.

Table 6
AUC and standard deviation for each model.

Model	AUC
RF	0.95 ± 0.021
GAM	0.95 ± 0.031
MaxEnt	0.94 ± 0.023
GLM	0.77 ± 0.032
ANN	0.79 ± 0.110

Although this case study is about birds, DeepData requires that the user inserts a depth so that later it can access the data. In this case, we consider a depth of 0, meaning that the birds are above the sea level. The paper uses AUC and True Skill Statistics as evaluation metrics. The use of the True Skill Statistics requires a threshold value, which is not specified in the study. Therefore DeepData only considers AUC.

Table 6 presents the AUC results and standard deviation for each model. As expected, both RF and GAM obtained high AUC values.

Since the ensemble model also considers Generalized boosted model that the tool cannot produce, the final distribution cannot be achieved. Fig. 7 presents the individual distributions for each model. The x and y axis in Fig. 7 represent latitude and longitude in the projected coordinate system. DeepData also generates model specific outputs, which are shown on Appendix B (Figs. 11 and 12).

The model specific outputs are partial dependence plots, which show the marginal effect that each feature has on the predicted outcome. These plots can be compared with the ones obtained by the paper although some precautions are necessary since, contrary to the ones on the original paper, these do not represent probabilities.

5. Conclusions and future work

In this work, we present a web-based machine learning tool named DeepData that provides a simple and efficient way for creating species distribution models, taking into account the user domain knowledge and allowing to experiment different variable combinations and different models, while turning it more efficient as the user does not have to think about programming. DeepData supports options for all parts of the modelling process: (i) data pre-processing, (ii) model selection and (iii) model evaluation. Furthermore, it enables to load data of species and environmental data concerning the Azores Exclusive

Economic Zone. Furthermore, the user can upload its own data of both species and environmental variables.

The developed tool provides a comprehensive user interface to perform the entire modelling process using different state-of-the-art approaches. Nowadays, the two most used software tools for SDM modelling are MaxEnt and R (Meineri, Deville, Grémillet, Gauthier-Clerc, & Béchet, 2015). The approaches used by these tools are quite different. While MaxEnt uses a click approach, R uses a syntax driven approach. Our tool is the balance between click and syntax driven approaches. By displaying the available options, the user clicks on the desired options and the tool generates the syntax for the R software. One concern with the click approach is that it works like a 'black-box' software, meaning that the details are hidden from the user. The paper provides a full specification of all default options and options available for all its processes, so that the user is fully conscious of the model.

Although the supported user interface is mostly composed of click options, flexibility is not compromised. Each modelling phase has its own options thus supporting tuning, while making it easier for inexperienced users. It also supports multiple SDMs to be fitted and compared simultaneously. This makes comparison between different models possible because both pre-processing and evaluation methods that are applied are the same.

The major limitation of the `DeepData` tool is that it does not take into account that some users might want to use data that is private. Most of the studies use data that belongs to the government and therefore data that is not for public use. One way to solve this issue is to provide information access control. Information access control is composed of authentication and authorization. Authentication is concerned with confirming that the user is who she/he says, while authorization is concerned with the level of access each user is granted.

CRediT authorship contribution statement

Leonor Oliveira e Silva: Software, Visualization, Investigation, Writing – original draft, Writing – review & editing. **Magda Resende:** Data curation, Software. **Helena Galhardas:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Vasco Manquinho:** Resources, Funding acquisition, Writing – original draft, Writing – review & editing. **Inês Lynce:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – original draft, 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.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia, Portugal (FCT) under projects UIDB/50021/2020, CMU/AIR/0022/2017 and DSAIPA/AI/0044/2018.

Appendix A. Response curves for the crabeater seals case study

See Figs. 8–10.

Appendix B. Partial dependence plot for the Trindade Petrel case study

See Figs. 11 and 12.

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. http://dx.doi.org/10.1111/j.1365-2664.2006.01214.x.

BIOMOD modeling options. (2016). https://www.rdocumentation.org/packages/biomod2/versions/3.3-7.1/topics/BIOMOD.

Brandt, A., Griffiths, H., Gutt, J., Linse, K., Schiaparelli, S., Ballerini, T., et al. (2014). Challenges of deep-sea biodiversity assessments in the Southern ocean. *Advances in Polar Science*, *25*, 204–212. http://dx.doi.org/10.13679/j.advps.2014.3.00204.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Carsten F. Dormann, M. B. A. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *10*(5), 609–628. http://dx.doi.org/10.1111/j.2007.0906-7590.05171.x.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://dx.doi.org/10.1007/BF00994018.

Costello, M., Basher, Z., Sayre, R., Breyer, S., & Wright, D. (2018). Stratifying ocean sampling globally and with depth to account for environmental variability. *Scientific Reports*, *8*, http://dx.doi.org/10.1038/s41598-018-29419-1.

Crase, B., Liedloff, A. C., & Wintle, B. A. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, *35*(10), 879–888. http://dx.doi.org/10.1111/j.1600-0587.2011.07138.x.

Dobson, A. J. (2010). *An introduction to generalized linear models* (2nd ed.). Taylor & Francis.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. http://dx.doi.org/10.1111/j.1600-0587.2012.07348.x.

Elith, J., & Leathwick, J. (2009a). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, *40*, 677–697. http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith, J., & Leathwick, J. R. (2009b). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435. http://dx.doi.org/10.1111/ele.12189.

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

Hastie, T., & Tibshirani, R. (2005). Generalized additive model. In *Encyclopedia of biostatistics*. American Cancer Society, http://dx.doi.org/10.1002/0470011815.b2a09018.

Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, *45*(5), 1372–1381. http://dx.doi.org/10.1111/j.1365-2664.2008.01524.x.

IOC-UNESCO and UNEP (2016a). *Large marine ecosystems: status and trends*. United Nations Environment Programme(UNEP).

IOC-UNESCO and UNEP (2016b). *The open ocean: status and trends*. United Nations Environment Programme(UNEP).

Iyer, C. V. K., Hou, F., Wang, H., Wang, Y., Oh, K., Ganguli, S., et al. (2021). Trinity: A no-code AI platform for complex spatial datasets. In D. D. Lunga, H. L. Yang, S. Gao, B. Martins, Y. Hu, X. Deng, S. D. Newsam (Eds.), *GeoAI@SIGSPATIAL 2021: proceedings of the 4th ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery* (pp. 33–42). ACM, http://dx.doi.org/10.1145/3486635.3491072.

Jiménez-Valverde, A. (2014). Threshold-dependence as a desirable attribute for discrimination assessment: Implications for the evaluation of species distribution models. *Biodiversity and Conservation*, *23*(2), 369–385. http://dx.doi.org/10.1007/s10531-013-0606-1.

Kölzsch, A., Davidson, S. C., Gauggel, D., Hahn, C., Hirt, J., Kays, R., et al. (2022). MoveApps - A serverless no-code analysis platform for animal tracking data. http://dx.doi.org/10.1101/2022.02.15.480513, BioRxiv, Cold Spring Harbor Laboratory.

Koundouri, P., & Giannouli, A. (2015). Blue growth and economics. *Frontiers in Marine Science*, *2*, 94. http://dx.doi.org/10.3389/fmars.2015.00094.

Krüger, L. (2018). Figures, model results and environmental data for population estimates of the Trindade Petrel (Pterodroma arminjoniana), Trindade Island. http://dx.doi.org/10.1594/PANGAEA.889927, PANGAEA;
Krüger, L. (2018). Population estimates of trindade petrel *(Pterodroma arminjoniana)* by ensemble nesting habitat modelling. *International Journal of Environmental Sciences & Natural Resources*, *10*(4), http://dx.doi.org/10.19080/IJESNR.2018.10.555793, Supplement to, https://juniperpublishers.com/ijesnr/pdf/IJESNR.MS.ID.555793.pdf.

Krüger, L. (2018). Population estimates of Trindade Petrel *(Pterodroma arminjoniana)* by ensemble nesting habitat modelling. *International Journal of Environmental Sciences & Natural Resources*, *10*(4), 1–13. http://dx.doi.org/10.19080/IJESNR.2018.10.555793.

Krüger, L., Paiva, V. H., Petry, M. V., Montone, R. C., & Ramos, J. A. (2018). Population estimate of Trindade Petrel *Pterodroma arminjoniana* by the use of predictive nest habitat modelling. *Bird Conservation International*, *28*(2), 197–207. http://dx.doi.org/10.1017/S0959270916000289.

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1), Article e02567. http://dx.doi.org/10.1002/ecs2.2567.

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, *34*(2), 232–243. http://dx.doi.org/10.1111/j.1600-0587.2010.06354.x.

Manel, S., Williams, H. C., & Ormerod, S. (2001). Evaluating presence–absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology*, *38*(5), 921–931. http://dx.doi.org/10.1046/j.1365-2664.2001.00647.x.

Meineri, E., Deville, A.-S., Grémillet, D., Gauthier-Clerc, M., & Béchet, A. (2015). Combining correlative and mechanistic habitat suitability models to improve ecological compensation. *Biological Reviews*, *90*(1), 314–329. http://dx.doi.org/10.1111/brv.12111.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23. http://dx.doi.org/10.2307/2332142.

Nachtsheim, D. A., Jerosch, K., Hagen, W., Plötz, J., & Bornemann, H. (2015). Crabeater seals *(Lobodon carcinophaga)* in the weddell sea during DRE1998 campaign, with link to files of gridded seal locations and environmental parameters for maxent analyses. http://dx.doi.org/10.1594/PANGAEA.855006, PANGAEA;
Nachtsheim, DA, et al. (2016). Habitat modelling of crabeater seals *(Lobodon carcinophaga)* in the weddell sea using the multivariate approach maxent. *Polar Biology*, *40*(5), 961-976. http://dx.doi.org/10.1007/s00300-016-2020-0, In supplement to.

Nachtsheim, D. A., Jerosch, K., Hagen, W., Plötz, J., & Bornemann, H. (2017). Habitat modelling of crabeater seals *(Lobodon carcinophaga)* in the weddell sea using the multivariate approach Maxent. *Polar Biology*, *40*(5), 961–976. http://dx.doi.org/10.1007/s00300-016-2020-0.

Palm, G. (1986). Warren McCulloch and walter pitts: A logical calculus of the ideas immanent in nervous activity. In G. Palm, & A. Aertsen (Eds.), *Brain theory* (pp. 229–230). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-70911-1_14.

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, *133*(3), 225–245. http://dx.doi.org/10.1016/S0304-3800(00)00322-7.

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *ICML: vol. 04*, *Proceedings of the twenty-first international conference on machine learning* (pp. 83–90). New York, NY, USA: ACM, http://dx.doi.org/10.1145/1015330.1015412.

Porfirio, L. L., Harris, R. M. B., Lefroy, E. C., Hugh, S., Gould, S. F., Lee, G., et al. (2014). Improving the use of species distribution models in conservation planning and management under climate change. *PLoS One*, *9*(11), 1–21. http://dx.doi.org/10.1371/journal.pone.0113749.

Schötteler, S., Laumer, S., Schuhbauer, H., Scheidthauer, N., Seeberger, P., & Miethsam, B. (2021). A no-code platform for tie prediction analysis in social media networks. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds.), *Innovation through information systems* (pp. 475–491). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-86797-3_32.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms* (1st ed.). Chapman & Hall/CRC.