Title

flexsdm: An r package for supporting a comprehensive and flexible species distribution modelling workflow

Journal

Methods in Ecology and Evolution, 13(8)

ISSN

2041-210X

Authors

Velazco, Santiago José Elías
Rose, Miranda Brooke
Andrade, André Felipe Alves
et al.

Publication Date

2022-08-01

DOI

10.1111/2041-210x.13874

Peer reviewed

Received: 13 September 2021 | Accepted: 7 April 2022

DOI: 10.1111/2041-210X.13874

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

FLEXSDM: An R package for supporting a comprehensive and flexible species distribution modelling workflow

Santiago José Elías Velazco[1,2] | Miranda Brooke Rose[1] | André Felipe Alves de Andrade[3] | Ignacio Minoli[4] | Janet Franklin[1]

[1]Department of Botany and Plant Sciences, University of California—Riverside, CA, USA

[2]Programa de Pós-Graduação em Biodiversidade Neotropical, Universidade Federal da Integração Latino-Americana, Foz do Iguaçu, Brazil

[3]Theory, Metacommunity and Landscape Ecology Lab, ICB V, Universidade Federal de Goiás, Goiânia, Brazil

[4]Instituto de Biología Subtropical, Universidad Nacional de Misiones—CONICET, Puerto Iguazú, Argentina

Correspondence
Santiago José Elías Velazco
Email: sjevelazco@gmail.com

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 165174/2020-0; National Science Foundation, Grant/Award Number: 1853697

Handling Editor: Huijie Qiao

Abstract

1. Species distribution models (SDM) are widely used in diverse research areas because of their simple data requirements and application versatility. However, SDM outcomes are sensitive to data input and methodological choices. Such sensitivity and diverse applications mean that flexibility is necessary to create SDMs with tailored protocols for a given set of data and model use.
2. We introduce the R package FLEXSDM for supporting flexible species distribution modelling workflows. FLEXSDM functions and their arguments serve as building blocks to construct a specific modelling protocol for user's needs. The main FLEXSDM features are modelling flexibility, integration with other modelling tools, simplicity of the objects returned and function speed. As an illustration, we used FLEXSDM to define a complete workflow for California red fir *Abies magnifica*.
3. This package provides modelling flexibility by incorporating comprehensive tools structured in three steps: (a) The Pre-modelling functions that prepare input, for example, sampling bias correction, sampling pseudo-absences and background points, data partitioning, and reducing collinearity in predictors. (b) The Modelling functions allow fitting and evaluating different modelling approaches, including individual algorithms, tuned models, ensembles of small models and ensemble models. (c) The Post-modelling functions include tools related to models' predictions, interpolation and overprediction correction.
4. Because FLEXSDM comprises a large part of the SDM process, from outlier detection to overprediction correction, FLEXSDM users can delineate partial or complete workflows based on the combination functions to meet specific modelling needs.

KEYWORDS
ecoinformatics, ecological niche modelling, ensemble modelling, model fit for purpose, model tuning, spatial ecology, spatially structured validation

1 | INTRODUCTION

Species distribution modelling (SDM) is a modelling technique that predicts species distribution range combining, in its simplest form, species occurrences and spatial environmental data (Franklin, 2010; Peterson et al., 2011). SDMs are capable of predicting different distribution types (potential vs. realized distribution Peterson et al., 2011), for different regions or periods (past or future), being combined with other biodiversity data (e.g. functional and phylogenetic diversity), or coupled with other modelling approaches (e.g. metapopulation or individual-based models). This simplicity in data needed and application versatility has led SDM to become a popular tool in research areas such as ecology, conservation, biogeography, palaeobiogeography and epidemiology (Franklin, 2013; Peterson et al., 2011).

SDM outcomes are sensitive to data input and methodological choices. For instance, suitability predicted by SDM can be affected by the area used for model calibration, type of data partitioning for calibration and validation, number and type of algorithms and their parametrization, or the number, quality, and spatial configuration of species records and absences, pseudo-absence, or background points. Such sensitivity and multiplicity of uses mean that flexibility is needed to create SDMs with tailored protocols for a given set of data and specific model use.

Here we present FLEXSDM, a new R package created to provide SDM workflow flexibility by incorporating comprehensive tools for data preparation, model fitting, prediction and evaluation (Franklin, 2010). FLEXSDM consists of >40 functions that can be used independently or combined with other modelling tools (Figure 1). One of the most exciting features of FLEXSDM is its broad parametrization capacity using a wide variety of functions and arguments. This allows users to define their complete or partial modelling procedure.

2 | PACKAGE OVERVIEW

The main features of FLEXSDM are as follows:

1. Modelling flexibility: One of the most important goals of this R package is to offer a flexible modelling procedure by combining many independent functions with multiple arguments that serve as building blocks to construct a specific modelling protocol for user's needs. Furthermore, dividing the modelling process into individual functions allows control and inspection of the output of each step (from filtering occurrence data to correcting overprediction).
2. Integration with other modelling tools: This modelling procedure is based on modular functions and allows users to use a required function without depending on this package to complete the modelling process. For instance, the outputs of the functions for defining calibration area, data partitions or a pseudo-absence sample can be saved and used with other modelling tools. Furthermore, data generated outside FLEXSDM can be used with FLEXSDM functions and workflows, such as a sampling bias layer for sampling background points or correcting the overprediction of models fitted using other software.
3. Simplicity of the object(s) returned by function output: FLEXSDM functions return a single or combination of four common R objects: (a) *tibble* tables from the DPLYR package, (b) *SpatialVect* vector spatial data and (c) *SpatRaster* raster spatial data, the last two from TERRA package, these three objects can be combined in a (iv) list class object. Additionally, Modelling functions return algorithm-specific objects (see Table 1). This feature facilitates a posteriori inspection, manipulation and exportation of function outputs, reducing memory requirements.
4. Function speed: Because FLEXSDM uses the recently developed TERRA package as a dependency, spatial data manipulation is faster and requires less computer memory than was used by the RASTER package. Additionally, FLEXSDM uses the FOREACH and DOPARALLEL packages, allowing parallel data processing and speeding up function execution.

In the FLEXSDM package, functions are grouped into three modelling steps. Pre-modelling includes functions that prepare input data (e.g. species occurrences thinning, sample pseudo-absences or background points, delimiting calibration area, reducing collinearity in predictors). The Modelling step includes functions related to model construction and validation. Post-modelling includes tools for geographical predictions, model evaluation and overprediction correction (Figure 1).

3 | PRE-MODELLING FUNCTIONS

3.1 | Delimitating a calibration area `calib_area`

Delimiting a calibration area (a.k.a. species' accessible area) is not a trivial decision because it interacts with other modelling steps, such as sampling pseudo-absences and background points, and affects performance metrics and predicted suitability patterns (Barve et al., 2011; Giovanelli et al., 2010; VanDerWal et al., 2009). The `calib_area` function delimits the calibration areas based on the following methods: 'buffer' defines a calibration area using a buffer around presences; 'mcp' uses a minimum convex polygon constructed on species presences; 'bmcp' uses a minimum convex polygon and adds a buffer around it; and 'mask' delimits the calibration area based on the polygons intersected by presences. The polygons used in the 'mask' method could be based on boundaries of ecological regions, vegetation types, hydrological basins or any other data that ecologically defines a species' accessible area. The package allows any externally defined calibration area to be named in the 'calibarea' argument found in `sample_background` and `sample_pseudoabs` function.

3.2 | Filtering occurrences—`occfilt_` function family

FLEXSDM provides functions for thinning species occurrences in geographical or environmental space implemented in the

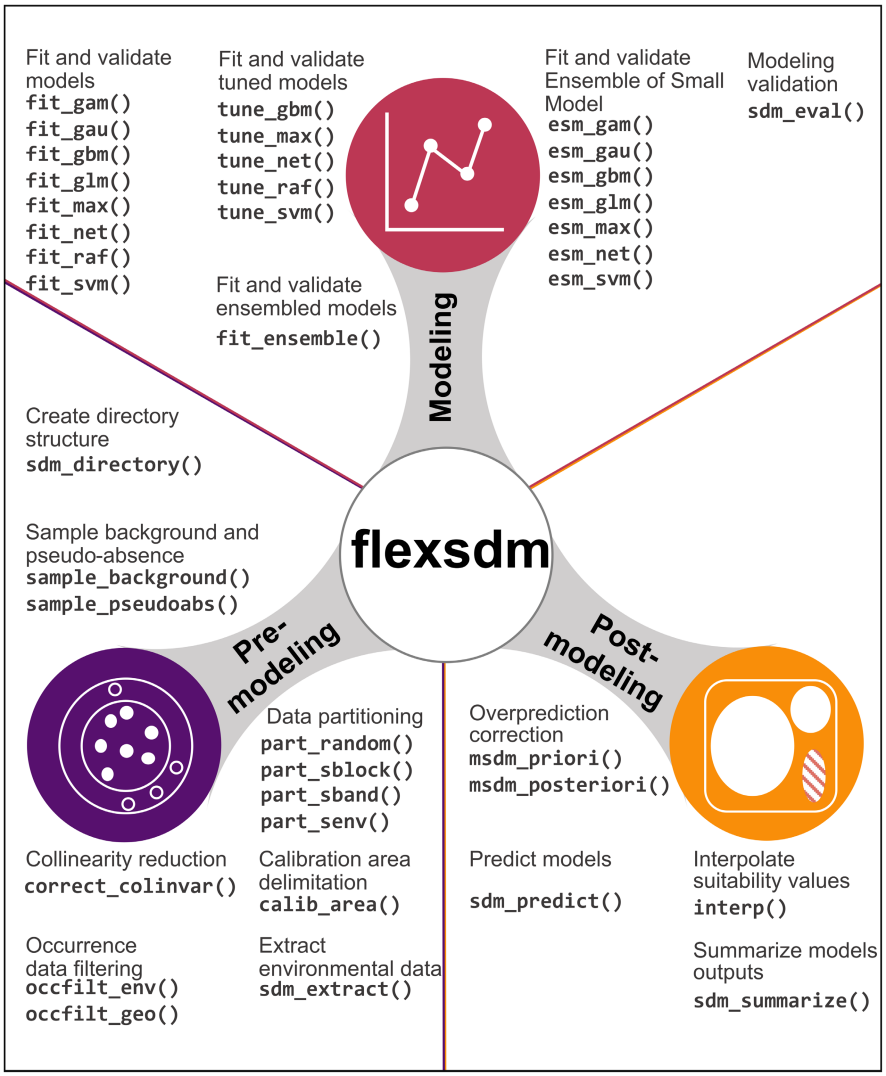

FIGURE 1 Functions of FLEXSDM package structured in three modelling steps: Pre-modelling (functions that prepare modelling input), modelling (functions related to model construction and validation) and post-modelling (tools related to models' predictions, inspection and correction)

TABLE 1 Algorithms implemented in FLEXSDM, the function family where were implemented, function suffix and package dependency

Algorithm name	Function family	Suffix	Package
Generalized Additive Models	fit_, tune_, esm_	gam	MGCV
Gaussian Process models	fit_, esm_	gau	GRAF
Generalized Boosted Regression	fit_, tune_, esm_	gbm	GBM
Generalized Linear Models	fit_, esm_	glm	STATS
Maximum Entropy	fit_, tune_, esm_	max	MAXNET
Artificial Neural Networks	fit_, tune_, esm_	net	NNET
Random Forest	fit_, tune_	raf	RANDOMFOREST
Support Vector Machine	fit_, tune_, esm_	svm	KERNLAB

occfilt_geo and occfilt_env functions. occfilt_geo uses the same process implemented in the R package SPTHIN to geographically filter occurrence points is based on a minimum nearest-neighbour distance (Aiello-Lammens et al., 2015). Three methods can be used to define minimum distance: 'moran' determines the threshold as the minimum distance which minimizes the spatial autocorrelation in occurrence data, following a Moran's semi-variogram; 'cellsize' filters occurrences based on the spatial resolution of predictors, and in addition that minimum distance can be adjusted to represent a larger grid cell size that are multiples of the original predictors; finally, 'determined' requires a user-defined minimum distance in kilometres. The environmental thinning function is based on Varela et al. (2014). However, occfilt_env has the advantage of running filtering procedure quickly, working with the original variables without performing a PCA beforehand, using any number of variables and multidimensional grid size. As far as

we know, this procedure with such improvements is not available in another R package.

3.3 | Pseudo-absence and background point sampling—`sample_` function family

FLEXSDM allows users to sample background points and pseudo-absences with the `sample_background` and `sample_pseudoabs` functions. We implemented three background allocation methods: 'random' randomly samples background points in a given study area; 'thickening' is a new method proposed by Vollering et al. (2019) that selects background points that are geographically biased towards species presences based on the superposition of buffers around presences; and 'biased' samples background points using a raster layer representing sampling effort of occurrences throughout the study area approach (e.g. presence kernel density, Fourcade et al., 2014; Phillips et al., 2009) without biasing towards presences as in 'thickening'. These methods allow modellers to generate background samples for specific situations (e.g. rare species or biased samples; Elith et al., 2010; Vollering et al., 2019).

Pseudo-absences sample size and allocation method affect SDM outcomes (Liu et al., 2019). The `sample_pseudoabs` function implements five allocation approaches: 'random' randomly allocates pseudo-absences throughout the area used for model fitting. 'env_const' environmentally constrains pseudo-absences to regions with lower suitability values predicted by a Bioclim model (Busby, 1991). With 'geo_const' pseudo-absences are sampled a given distance away from presences. 'geo_env_const' combines 'env_const' and 'geo_const'. 'geo_env_km_const' constrains sampling using a three-level procedure, similar to geo_env_const with an additional step that distributes the pseudo-absences in environmental space using a *K*-means cluster analysis.

Both functions can sample data within a calibration area or regions with given values (e.g. groups delimited by geographical block or band partition). Spatial constraint types are relevant because they control how much data will be allocated in a defined region (see example, below).

3.4 | Data partitioning—`part_` function family

We implemented a broad range of data partition methods suitable for different modelling conditions, such as the amount of data, modelling approach or use of models. `part_random` performs partitioning based on the random selection of training and testing groups, such as k-fold, repeated k-fold, leave-one-out cross-validation (a.k.a. Jackknife) and bootstrapping (Fielding & Bell, 1997). The other `part_` functions are geographically or environmentally structured. Structured partition methods are required to evaluate model transferability, relevant when using SDMs to project models onto different periods or regions (Roberts et al., 2017; Santini et al., 2021).

The `part_sband` and `part_sblock` functions allow testing with different numbers of partitions using latitudinal or longitudinal bands, or square blocks, respectively. Both functions explore a range of band or block sizes and automatically select the best size for a given presence, presence–absence or presence-pseudo-absence dataset. Size selection uses an optimization procedure that explores partition size in three dimensions determined by spatial autocorrelation (measured by Moran's *I*), environmental similarity (Euclidean distance) and differences in the amount of data among partition groups (Standard Deviation—*SD*; Velazco et al., 2019). This procedure will iteratively select those partitions with autocorrelation values less than the first quartile of Moran's *I*, then those with environmental similarity values greater than the third quartile of the Euclidean distances, and finally, those with a difference in the amount of data less than the first quartile of *SD*. This selection is repeated until only one partition is retained (Velazco et al., 2019). The optimization procedure uses quartiles because it pragmatically selects the best subset for each parameter, quartiles being more restrictive than just using the mean. This partition selection method: (a) is not subjective, (b) balances the environmental similarity and spatial autocorrelation between partitions, and (c) controls the selection of partitions with few data that may be problematic for model fitting.

`part_senv` partitions data based on its environmental condition. This function explores a broad range of partition numbers (i.e. groups in environmental space) based on *K*-mean clustering and returns the number best suited for a given presence, presence–absence or presence-pseudo-absence database. `part_senv` selects the best partition number using the same optimization procedure used in `part_sband` and `part_sblock`.

All `part_` functions will add columns to an occurrence dataset with partition groups (using column names that start with '.part'). Because partitions are stored in columns, it is possible to use partition groups defined outside of the package, for instance, partitioning using temporal bins or native vs. invaded localities. Because these partition groups are stored as numeric or text (using the words 'train' and 'test') and with column names that start with .part, the Modelling functions will interpret them as partition information.

3.5 | Collinearity correction of predictors variable `correct_colinvar`

Predictor collinearity is a common issue for species distribution models, leading to model overfitting and unstable parameters estimation, which affect model projections (Brun et al., 2020; De Marco & Nóbrega, 2018). `correct_colinvar` has four methods to deal with collinearity: 'pearson' detects pair of predictors with a Pearson correlation index higher than a determined threshold; 'vif' is based on the variance inflation factor (VIF) and removes those predictors with a higher VIF than the chosen threshold; 'fa' performs a factorial analysis to reduce dimensionality and selects the predictor with

the highest correlation to each axis; finally, 'pca' performs a principal components analysis in the predictors and returns the axes that account for 95% of the total variance in the system as predictors.

3.6 | Additional pre-modelling functions

The `sdm_directory` function assists in organizing inputs (e.g. occurrences, predictors) and outputs by creating folders (directories) based on user specifications, such as choice of algorithms, ensemble methods and model projections to new geographical regions or periods. To implement SDMs in FLEXSDM, users first need a database with spatial coordinates of species occurrences, absences/background points/pseudo-absences and environmental data extracted at each point location. The `sdm_extract` function returns the original database and additional columns of the environmental values stored in the raster object.

4 | MODELLING FUNCTIONS

Because the choice of modelling algorithm is one of the main sources of model uncertainty (Watling et al., 2015) and there is no single modelling method that can perform well in all modelling situations (Qiao et al., 2015), FLEXSDM allows construction of models based on different algorithms and approaches (Table 1). Thus, we developed the following function families: (a) `fit_`: creates models based on default hyperparameter values or with user-specified values; (b) `tune_`: creates models by hyperparameter tuning, (c) `esm_`: uses the ensembles of small models approach (Breiner et al., 2015) and (d) `fit_ensemble`: create ensemble models. All these functions work with continuous and categorical predictors and allow parametrization with user-specified formulas. For `fit_`, `tune_` and `esm_` function families, an algorithm name is specified with the last three letters of the functions (Table 1). All Modelling functions create and validate models with any combination of outputs from the Pre-modelling functions.

4.1 | `fit_` family functions

This family of functions fits and validates eight algorithms with default hyperparameters values. However, hyperparameters are provided as arguments so that the user can specify other values.

4.2 | `tune_` family functions

Models' hyperparameter values affect the degree of model complexity and the predicted species' geographical range; therefore, selecting the best hyperparameter values for a given dataset could be preferred over default ones (Fourcade, 2021; Morales et al., 2017). The `tune_` family functions allow tuning of eight algorithms based on a grid search approach, that is, 'brute-forcing' all possible combinations of hyperparameter values. One of the most important features of tuning in FLEXSDM is that it can be based on any partition methods, thresholds and performance metrics. For instance, it is possible to tune a Maximum Entropy model based on spatial block partitioning, the Boyce metric and one or more thresholds. We recommend tuning models whenever possible because the best values of hyperparameters are specific to a modelling condition (e.g. partition method, number of records, predictors variables). However, the time for running tuned models can be considerably higher than those using default hyperparameters.

4.3 | `esm_` family functions

Ensembles of small models is an approach for modelling rare or poorly sampled species (Breiner et al., 2015). This method creates bivariate (two predictors) models with all the pairwise combinations of available predictors and performs an ensemble based on the average of suitability weighted by Somers' *D* metric (an example for modelling a rare species is available at https://sjevelazco.github.io/flexsdm/articles/v05_Rare_species_example.html

4.4 | `fit_ensemble` function

An ensemble approach in SDM is used to reduce model uncertainty or get a consensus prediction (Araújo & New, 2007). The function `fit_ensemble` allows users to perform and validate an ensemble based on different modelling methods. The ability to validate an ensemble model is a notable attribute of our package. The following ensemble methods are available: (a) average of different models' suitability; (b) weighted average based on model performance; (c) average of the best models, the set of algorithms with a higher-than-average performance among the algorithms (for this and previous methods, it is possible to use any performance metric); (d) averaging performed only on cells with suitability values above the selected threshold; and (e) median of models' suitability.

4.5 | Model performance metrics and thresholds—`sdm_eval`

FLEXSDM supports model evaluation based on several performance metrics and threshold types (Freeman & Moisen, 2008; Leroy et al., 2018) using the `sdm_eval` function, which interacts with all previous modelling functions. The range of performance metrics is important because some will be more appropriate than others depending on, for example, the quality of species occurrence data or data type used for modelling (Table 2; Leroy et al., 2018). Threshold criteria are used to divide SDM-generated probability values into a binary (present, absent) prediction, either for evaluation (Table 2) or for application of the SDM. The threshold criteria (Liu et al., 2005)

Performance metric	Dependent of threshold	Range
True-Positive Rate or Sensitivity (TPR)	Yes	0–1
True-Negative Rate or Specificity (TNR)	Yes	0–1
Sorensen	Yes	0–1
Jaccard	Yes	0–1
F-measure on presence-background (FPB)	Yes	0–2
Omission Rate (OR)	Yes	0–1
True Skill Statistic (TSS)	Yes	−1–1
Kappa	Yes	0–1
Area Under Curve (AUC)	No	0–1
Continuous Boyce index (BOYCE)	No	−1–1
Inverse Mean Absolute Error (IMAE)[a]	No	0–1

[a]IMAE is calculated as 1 − (Mean Absolute Error) to be consistent with the other metrics where the higher the value of a given performance metric, the greater the model's accuracy.

TABLE 2 Performance metric implemented in FLEXSDM, their dependency of threshold and range

available are those at which: (a) there is no omission, (b) the sensitivity and specificity are equal, (c) the threshold that maximizes the TSS), (d) Jaccard is the highest, (e) Sorensen is highest, (f) FPB is highest, or (g) based on a specified sensitivity value.

5 | POST-MODELLING FUNCTIONS

5.1 | Method for correcting model overprediction

Sometimes, models are needed that estimate ranges close to realized distributions (Peterson & Soberón, 2012). Models that overpredict this range can produce misleading results and misguide conservation assessments (Velazco et al., 2020). FLEXSDM provides nine constraining methods adapted from the code of the MSDM package (Mendes et al., 2020). These constraining methods can be grouped into those that generate predictor variables that are used in modelling fitting (`msdm_priori`) and those where constraints are based on the interaction of observed species presences and suitability patterns (`msdm_posteriori`). Some of these methods can increase the error of omission; consequently, they must be used carefully with species with low detectability (Mendes et al., 2020). We advise using these approaches to create models used only for current conditions and not for different time periods (past or future).

5.2 | Additional post-modelling functions

`sdm_summarize` merges model performance tables from different modelling functions. `sdm_predict` performs spatial predictions for individual or ensemble models, producing maps of continuous suitability, binary predictions based on one or more specified thresholds or continuous suitability above a given threshold. `Interp` makes annual maps of suitability by interpolating between suitability maps for two time periods; interpolation could be useful when SDMs are coupled to, for example, spatially explicit population or metapopulation models.

5.3 | FLEXSDM and other SDM packages for R

Over the last decade, several R packages have been published offering complete or partial SDM workflows. We found that, on average, 24% of FLEXSDM features are shared with other R packages (Tables S1 and S2). The most similar packages were ENMTML (64%), BIOMOD2 (37%) and SSDM (32%). Features that distinguish FLEXSDM from most other packages (i.e. are unique to FLEXSDM or are shared with only a few other packages) are related to occurrences filtering, geographical and environmental partitioning, algorithm tuning based on any partition methods, some threshold types and performance metrics, ensembles of small models, and overprediction correction (Table S2). Although FLEXSDM shares several features with ENMTML (and not with other packages), the main difference is that FLEXSDM allows these features to be used independently of each other.

SDM is still a developing field in which new and improved methods are constantly being proposed. Consequently, it is difficult for a package to offer all modelling possibilities and always be up to date. FLEXSDM encompasses a broad range of the SDM workflow; however, other packages provide alternative methods and complementary tools. It is necessary to integrate different packages to implement state-of-the-art SDM. FLEXSDM supports such integration because modelling procedures are constructed with individual functions and simple and easily handled objects are returned. For instance, the pre-modelling functions can be used in combination with packages that lack pre-modelling functionality such as *biomod2*, *ecospat*, *ENMeval*, *kuenm*, *sdm* or *SDMtune*. Additionally, the overprediction correction and interpolation functions can be used with predicted distributions of any other package (Table S2).

6 | EXAMPLE

We illustrate the use of FLEXSDM to model the distribution of California red fir *Abies magnifica* (https://sjevelazco.github.io/flexsdm/articles/v04_Red_fir_example.html). Red fir is a high-elevation conifer tree

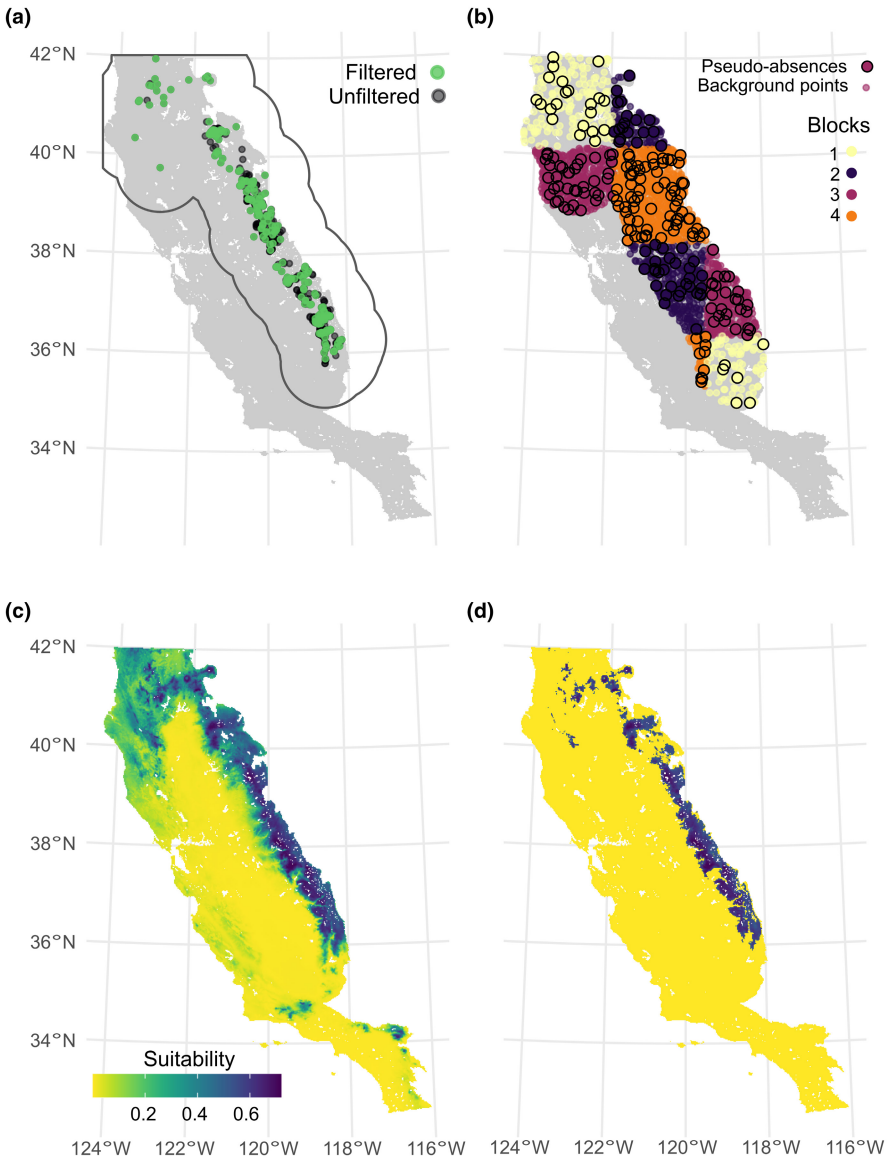

FIGURE 2 FLEXSDM outputs for the example modelling workflow for California red fir *Abies magnifica*. (a) Calibration area delimited by a buffer around unfiltered presences, then the sampling bias of presences was corrected using an environmental approach. (b) Pseudo-absences and background points sampled within each partition block, both weighted by the number of presences in each partition. (c) Suitability predicted by an ensemble model. (d) Suitability overprediction was corrected by occurrences based restriction method

species whose geographical range extends through the Sierra Nevada in California into the Cascade Ranges of Oregon, USA. We used species occurrence data compiled from several sources (Hannah et al., 2008; calflora.org; wildlife.ca.gov/Data/BIOS) with the following predictors: climatic water deficit, summer precipitation, winter precipitation, and minimum temperature of the coldest month sourced from the Basin Characterization Model (http://climate.calcommons.org/bcm).

A calibration area was delimited by a 100-km buffer around presences using `calib_area`; then, species presences were environmentally filtered using `occfilt_env` with eight bins (Figure 2a). With `part_sblock`, the presence data were spatially partitioned into four groups, after testing 30 grid-sizes. The `get_block` function was used to produce a map layer with partition information using the resolution and extent of the environmental variables. Then we used this layer for allocating a stratified sample of pseudo-absences and background points with `sample_background` and `sample_pseudoabs`, respectively. The number of pseudo-absences and background points were set equal to 10 times the number of presences found in each partition, respectively (Figure 2b). We used the modelling algorithms Maxent, Gaussian Process and Generalized Linear Model, which were fitted and validated by `tune_max`, `fit_gau` and `fit_glm`. The final model consisted of a weighted mean ensemble fitted and validated with `fit_ensemble` and predicted with `sdm_predict` (Figure 2c). Finally, we used the occurrences-based restriction method (Mendes et al., 2020) in the `msdm_posteriori` to correct model overprediction (Figure 2d).

7 | CONCLUSIONS

FLEXSDM is a new R package that offers comprehensive and flexible tools for species distribution modelling, ranging from outlier detection to overprediction correction. FLEXSDM users can delineate partial or complete modelling workflows based on the combination of >40 functions to meet specific modelling needs. The main FLEXSDM features are its modelling flexibility, integration with other modelling

tools, simplicity of the objects returned, and function speed. Novel innovations in this SDM package include model tuning functions based on any partition methods, thresholds, and performance metrics, ensemble model validation, flexible data partitioning, environmental occurrence filtering, and overprediction correction.

AUTHORS' CONTRIBUTIONS

S.J.E.V., A.F.A.d.A., M.B.R., I.M. and J.F. wrote the R package; M.B.R., I.M. and J.F. wrote the vignette; S.J.E.V., M.B.R. and I.M. conducted the package tests; all authors wrote the package website and function documentations; S.J.E.V. led the manuscript writing, and all authors wrote the manuscript.

ACKNOWLEDGEMENTS

S.J.E.V., M.B.R. and J.F. were supported by the National Science Foundation, USA (Award 1853697 to H.M Regan and J. Franklin). A.F.A.d.A. is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - 165174/2020-0). We thank J. H. Thorne, University of California Davis, and R. Yacoub, California Dept. Fish & Wildlife, for access to species occurrence data used in the example.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13874.

DATA AND PACKAGE AVAILABILITY STATEMENT

Codes and data used in this manuscript, package documentations and vignettes are available at GitHub (https://sjevelazco.github.io/flexsdm) and Zenodo (https://doi.org/10.5281/zenodo.6462044, Velazco et al., 2022) repositories. The development version of FLEXSDM is https://github.com/sjevelazco/flexsdm.

ORCID

Santiago José Elías Velazco https://orcid.org/0000-0002-7527-0967
Miranda Brooke Rose https://orcid.org/0000-0001-5269-3759
André Felipe Alves de Andrade https://orcid.org/0000-0002-6134-3176
Ignacio Minoli https://orcid.org/0000-0002-6039-6962
Janet Franklin https://orcid.org/0000-0003-0314-4598

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541–545. https://doi.org/10.1111/ecog.01132

Araújo, M., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*(11), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, *6*(10), 1210–1218. https://doi.org/10.1111/2041-210X.12403

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., & Zimmermann, N. E. (2020). Model complexity affects species distribution projections under climate change. *Journal of Biogeography*, *47*(1), 130–142. https://doi.org/10.1111/jbi.13734

Busby, J. R. (1991). BIOCLIM—A bioclimate analysis and prediction system. In C. R. Margules & M. P. Austin (Eds.), *Nature conservation: Cost effective biological surveys and data analysis* (pp. 64–68). CSIRO.

De Marco, P., & Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS ONE*, *13*(9), e0202403. https://doi.org/10.1371/journal.pone.0202403

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species: The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecological Modelling*, *457*, 109686. https://doi.org/10.1016/j.ecolmodel.2021.109686

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS ONE*, *9*(5), e97122. https://doi.org/10.1371/journal.pone.0097122

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

Franklin, J. (2013). Species distribution models in conservation biogeography: Developments and challenges. *Diversity and Distributions*, *19*(10), 1217–1223. https://doi.org/10.1111/ddi.12125

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, *217*(1), 48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015

Giovanelli, J. G., de Siqueira, M. F., Haddad, C. F., & Alexandrino, J. (2010). Modeling a spatially restricted distribution in the neotropics: How the size of calibration area affects the performance of five presence-only methods. *Ecological Modelling*, *221*(2), 215–224.

Hannah, L., Midgley, G., Davies, I., Davis, F., Ries, L., Thuiller, W., Thorne, J., Seo, C., Stoms, D., & Snider, N. (2008). BioMove—Improvement and parameterization of a hybrid model for the assessment of climate change impacts on the vegetation of California. California Energy Commission, Public Interest Energy Research Program, 91.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*, 1994–2002. https://doi.org/10.1111/jbi.13402

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385–393.

Liu, C., Newell, G., & White, M. (2019). The effect of sample size on the accuracy of species distribution models: Considering both presences and pseudo-absences or background sites. *Ecography*, *42*(3), 535–548. https://doi.org/10.1111/ecog.03188

Mendes, P., Velazco, S. J. E., de Andrade, A. F. A., & De Marco, P. (2020). Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy. *Ecological Modelling*, *431*, 109180. https://doi.org/10.1016/j.ecolmodel.2020.109180

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

Peterson, A. T., & Soberón, J. (2012). Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza & Conservação*, *10*(2), 102–107. https://doi.org/10.4322/natcon.2012.019

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton University Press.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. https://doi.org/10.1111/2041-210X.12397

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., & Huijbregts, M. A. J. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, *27*, 1035–1050. https://doi.org/10.1111/ddi.13252

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*(4), 589–594. https://doi.org/10.1016/j.ecolmodel.2008.11.010

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*, 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Velazco, S. J. E., Ribeiro, B. R., Laureto, L. M. O., & De Marco Júnior, P. (2020). Overprediction of species distribution models in conservation planning: A still neglected issue with strong effects. *Biological Conservation*, *252*, 108822. https://doi.org/10.1016/j.biocon.2020.108822

Velazco, S. J. E., Rose, M. B., de Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). FLEXSDM: An R package for supporting a comprehensive and flexible species distribution modeling workflow (V1.3.0). *Zenodo*, https://doi.org/10.5281/zenodo.6462044

Velazco, S. J. E., Villalobos, F., Galvão, F., & De Marco Júnior, P. (2019). A dark scenario for Cerrado plant species: Effects of future climate, land use and protected areas ineffectiveness. *Diversity and Distributions*, *25*(4), 660–673. https://doi.org/10.1111/ddi.12886

Vollering, J., Halvorsen, R., Auestad, I., & Rydgren, K. (2019). Bunching up the background betters bias in species distribution models. *Ecography*, *42*(10), 1717–1727. https://doi.org/10.1111/ecog.04503

Watling, J. I., Brandt, L. A., Bucklin, D. N., Fujisaki, I., Mazzotti, F. J., Romañach, S. S., & Speroterra, C. (2015). Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, *309–310*, 48–59. https://doi.org/10.1016/j.ecolmodel.2015.03.017

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Velazco, S. J., Rose, M. B., de Andrade, A. F., Minoli, I. & Franklin, J. (2022). flexsdm: An r package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution*, *00*, 1–9. https://doi.org/10.1111/2041-210X.13874