Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods

PRESENCE-ONLY AND PRESENCE-ABSENCE DATA FOR COMPARING SPECIES DISTRIBUTION MODELING METHODS

Jane Elith[1*], Catherine H. Graham[2], Roozbeh Valavi[1], Meinrad Abegg[2], Caroline Bruce[3], Simon Ferrier[4], Andrew Ford[5], Antoine Guisan[6], Robert J. Hijmans[7], Falk Huettmann[8], Lucia Lohmann[9], Bette Loiselle[10], Craig Moritz[11], Jake Overton[12], A. Townsend Peterson[13], Steven Phillips[14], Karen Richardson[15], Stephen E. Williams[16], Susan K. Wiser[17], Thomas Wohlgemuth[2], Niklaus E. Zimmermann[2]

[1]School of BioSciences, University of Melbourne, Australia. [2]Swiss Federal Research Institute WSL, CH-8903 Birmensdorf, Switzerland. [3]CSIRO Land and Water, Cairns, Queensland, Australia. [4]CSIRO Land and Water, Canberra, Australian Capital Territory (ACT), Australia. [5]CSIRO Land and Water, Tropical Forest Research Centre, Atherton, Queensland, Australia. [6]University of Lausanne, 1015 Lausanne, Switzerland. [7]University of California, Davis, USA. [8]EWHALE Lab, Institute of Arctic Biology, Biology & Wildlife Department, University of Alaska Fairbanks, Fairbanks Alaska 99775 USA. [9]Universidade de São Paulo, Brazil. [10]College of Agricultural and Life Sciences, University of Florida, USA. [11]Research School of Biology & Center for Biodiversity Analysis, Australian National University, Australia. [12]Manaaki Whenua—Landcare Research, Hamilton, New Zealand (current address: PANTHERA, Floor 18, 8 West 40 St, New York, USA 10018. [13]Biodiversity Institute, University of Kansas, Lawrence, Kansas 66045, USA. [14]Center for Biodiversity and Conservation, American Museum of Natural History, New York, USA. [15]Department of Geography, Planning and Environment, Concordia University, Montreal, Canada. [16]Centre for Tropical Environmental and Sustainability Science, James Cook University, Townsville, Australia. [17]Manaaki Whenua—Landcare Research, Lincoln, New Zealand

**Corresponding Author: j.elith@unimelb.edu.au*

Abstract. Species distribution models (SDMs) are widely used to predict and study distributions of species. Many different modeling methods and associated algorithms are used and continue to emerge. It is important to understand how different approaches perform, particularly when applied to species occurrence records that were not gathered in structured surveys (e.g. opportunistic records). This need motivated a large-scale, collaborative effort, published in 2006, that aimed to create objective comparisons of algorithm performance. As a benchmark, and to facilitate future comparisons of approaches, here we publish that dataset: point location records for 226 anonymised species from six regions of the world, with accompanying predictor variables in raster (grid) and point formats. A particularly interesting characteristic of this dataset is that independent presence-absence survey data are available for evaluation alongside the presence-only species occurrence data intended for modeling. The dataset is available on Open Science Framework and as an R package and can be used as a benchmark for modeling approaches and for testing new ways to evaluate the accuracy of SDMs.

From 2002 to 2005 a working group funded by the United States' National Center for Ecological Analysis and Synthesis (NCEAS), and led by ATP and CM, compared methods for fitting species distribution models (SDMs). These models combine observations of species occurrence or abundance with environmental data and can be used to predict distributions across space and time.

The authors of this current paper are the subset of the NCEAS working group who gathered and processed the data described here, alongside suppliers of those data; referred to here as "the NCEAS data group." The data come from six regions of the world (Fig. 1). For each region we gathered two types of species occurrence data: presence-only (PO, also known as "collection") data, and presence-absence (PA, or "survey") data. We generated random locations for each study region, referred to as "background" (or elsewhere "pseudo-absence") samples. These are necessary for model fitting for some modeling methods. We compiled spatially continuous environmental predictor variables ("raster data") deemed relevant to the species, and sampled these rasters at all PO, PA and background (BG) locations.

The NCEAS working group designed a "baseline" study to compare 16 modeling algorithms (Elith

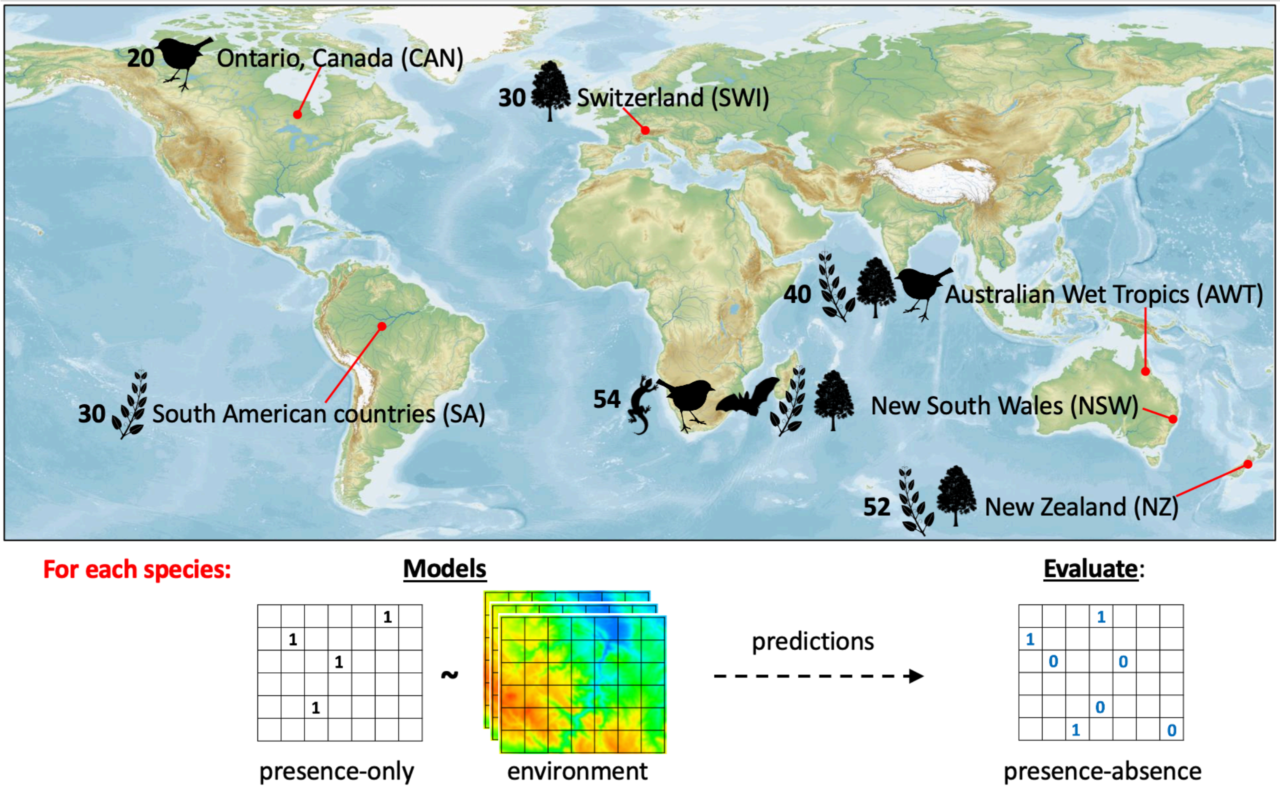


Figure 1: Overview of data supplied & model workflow. Numbers on map refer to number of species, and icons represent taxa (birds, trees, other plants, reptiles, bats), from each region indicated. The workflow at the bottom illustrates supply of presence-only species data with accompanying environmental covariates for modeling, and presence-absence (1/0) data at different sites, for evaluation.

et al. 2006), and also several experimental treatments that manipulated the datasets to explore the effects of sample size (Wisz et al. 2008), spatial resolution (grain) of environmental data (Guisan et al. 2007), error in PO location (Graham et al. 2008), bias in records (Dudik and Phillips 2009; Phillips et al. 2009) and treatment of BG data (Phillips et al. 2009) on model performance. Models were fitted (trained) on PO and optionally BG data. The environmental data for PA sites were provided, so modelers could predict environmental suitability for all species at these sites. In all studies, models were fitted by working group members with expertise in the respective methods, and modeling was blind to the species presence or absence at evaluation sites. The modelers sent their predictions to one group member who evaluated them against the PA observations at those sites. In subsequent years thirteen additional papers were published (detailed in Supplementary Information 1,[1] exploring aspects of the data or of model performance. The NCEAS data group obtained and compiled the data in 2002. In support of recent trends to make science more transparent and repeatable (National Academy of Sciences et al. 2009; Zuckerberg et al. 2010; Garzon-Lopez et al. 2016; Munafò et al. 2017) we have now obtained permissions to publish the data.

These data are valuable for their spread across regions of the world and across species, and particularly for the complementary sets of PO and PA species data. Here we expand on the latter point. Whilst SDMs can be fitted to a range of data types, the use of PO and presence-background (P-BG) are common (Kissling et al. 2018). Fit and evaluation of SDMs is often achieved through cross-validation; that is, using subsets of the data iteratively for either training and fitting the model or for testing (evaluation) (Hastie et al. 2009; Hijmans 2012; Roberts et al. 2017). A problem with PO data is that they can be spatially biased with some areas sampled intensively and others not at all (Reddy and Dávalos 2003; Hortal et al. 2008; Amano and Sutherland 2013; Isaac and Pocock 2015) and thus, may not be representative of the species distribution in the study area. When evaluating with PO or P-BG data, such biases remain, thus

[1] http://hdl.handle.net/1808/30579

wrongly emphasising the suitability of some environments and under-reporting the suitability of others. A model trained *and* evaluated with biased PO or P-BG data may appear to perform well, although it does not produce meaningful predictions of the true species distribution (El-Gabbas and Dormann 2018). Whilst methods are available for dealing with these sorts of bias in model training (Phillips et al. 2009; Syfert et al. 2013; Warton et al. 2013; Dorazio 2014; Fithian et al. 2015; Stolar and Nielsen 2015; Qiao et al. 2017), none are problem-free, and SDM evaluation remains a challenge. Although the spatial distribution of PA locations can also be biased, the bias is less problematic because a higher or lower spatial density of sites in some areas simply leads to a more or less precise undestanding of the species distribution (assuming that at least some sites exist across the major environmental gradients) (Phillips et al. 2009). PA evaluation data are therefore very useful. They can allow for an independent and less biased view of whether models correctly predict species occurrences. They also can be used to calculate a broader suite of evaluation statistics than those that can be estimated from PO data (Lawson et al. 2014). Whilst PA data are desirable for evaluation (El-Gabbas and Dormann 2018) they are often not available, so the data supplied here—with both PO and PA data from independent sources - are a valuable resource for testing new modeling methods and evaluation approaches.

Gathering and Processing the Data

Here we focus on the data used by the studies produced by the NCEAS working group (Supplementary Information 1). We are unable to release the original records supplied to us, and instead are releasing the cleaned version used in our modeling. We have permission to release anonymized species labels rather than real species names; this will not detract from the dataset as a benchmark resource for reproducing previous results or for assessing other aspects of fitting and evaluating SDMs. Some of the data preparation methods were reported in the original baseline modeling paper (Elith et al. 2006), but we describe them here in full detail, to gather all the information in one place, and to ensure the descriptions are adequate for data re-use. This manuscript and the accompanying metadata should be treated as the authoritative description of the data supplied here.

Data suppliers were initially asked to select species encompassing a range of life forms, responses to the environment, geographic distributions, and rarity, and to attempt to find species that had at least 20 records in both PO and PA datasets. The limit of 20 was set so we had enough information for training and testing models (Harrell 2001). This was generally adhered to, with some exceptions as evident in tables presented in Supplementary Information 2.[2] ("Summaries of species data for each region"). PO and PA datasets were to be from different collection efforts, and not have sites in common. Suppliers were asked to find a set of predictor variables in raster (grid) format that they considered relevant to the distribution of the species, and typical for what a skilled distribution modeler in their region would use. We asked for between ten and 15 variables to enable meaningful predictions and limit duplication across predictor variables, at the finest spatial resolution (smallest raster cell size) available, with a minimum acceptable resolution of 1 km^2. The minimum grain reflects the finest grain of global climate data available at that time.

All datasets were cleaned by JE and CG to these common properties agreed to by the group: (a) all data projected to a common projection for that region; (b) all raster data for a region aligned to the same extent and resolution, and only rasters with close to complete coverage in the region of interest retained; (c) species records reduced to a maximum of one record per raster cell using the following protocol: for PO data: if there is at least one presence record in a cell, retain one presence record for that cell; for PA data: reduce to one record per cell using the rule: if presence(s) and absence(s) both occur in the same cell, retain one presence; (d) records checked and rectified if necessary to ensure that PO and PA locations do not co-occur in a grid cell; (e) species records from locations with no environmental data removed.

Many SDMs contrast the environment at locations of known occurrence of a species to that at a set of random locations in the study region (background, quadrature, or pseudo-absence points: (Phillips et al. 2009; Warton and Shepherd 2010; Barbet-Massin et al. 2012; Renner et al. 2015). The NCEAS data group therefore supplied modelers with a sample of 10,000 background points for each region. Regional extents were delineated by the boundaries of countries or bioregions within countries, as deemed appropriate by the data suppliers. Background points were selected spatially at random across each region and sampled irrespective of the location of any presence records. That is, by design, a presence record and a

[2] http://hdl.handle.net/1808/30581.

background point might occur in the same raster cell. This approach aligns with recent interpretations of background samples as an approach for fitting a point process (Renner et al. 2015).

Characteristics of the Gathered Data

We sourced datasets from six regions of the world (Figure 1 and Table 1); the regions are hereafter referred to by the initials provided in Figure 1 and in column 1, Table 1. The six regions vary in size from approximately 24 to 12,223 thousand km^2 (Table 1). We gathered 11 to 13 predictor variables per region, most of which were continuous variables, but with four out of the six regions providing one or two categorical variables in their predictor sets (summarised in Table 1, and details of variables in Supplementary Information 3).[3] Records span species of birds (AWT, CAN, NSW), bats (NSW), plants (AWT, NSW, NZ, SA, SWI) and reptiles (NSW) (Table 1), totaling 226 species. The regions show useful variation in the amount of species data per region, with tens to thousands of PO records per species, and 102 to 19 120 PA evaluation sites (Table 1 and detailed summaries per region in Supplementary Information 2). This provided a diverse and representative data set for the NCEAS studies (Supplementary Information 1), and a benchmark set that we anticipate being broadly useful into the future.

Data sources for the PO and PA species data are detailed in Table 2. The different data sources used different sampling designs and methods which can provide insights into how data quality influences model outcomes/accuracy. For instance, some PO locations are recorded by GPS and therefore likely accurate (see AWT, Table 2), whereas others are typical PO data from a range of sources where the level of geographic accuracy is unknown (see NSW, Table 2). All the PA data are from intentional surveys, but vary in their design, age and number of data collectors. For instance, the SWI data are from plots on a regular lattice, the SA data are all collected by one person, and the CAN data are breeding bird data collected over years by multiple people. These variations are typical of what is seen in ecological datasets further making this dataset a useful benchmark for SDM modelers.

Details of Data Format and Location

The data are available on Open Science Framework (OSF) and accompanied by human metadata and/or readme files, and machine-accessible metadata; most data are also available in an R package, as described below. All data, organised as described below, are available openly.[4]

A. OSF data

In overview, we have uploaded the data in separate directories. The environmental raster (gridded) data are separate from all other records, since their zip file is 561 MB in total and many users will not want to predict to the rasters but rather to the tabled environmental data for the evaluation sites. The species data (both PO and PA), and the background samples are in separate folders within a “Records” folder and include the data extracted from these rasters (i.e. all environmental conditions) at every site, and thus are ready for modeling and evaluation. All site-based data (PO, BG, PA) are available as comma-separated text files (.csv). Polygon outlines of each region are also supplied to give context to locations of species records, if users want to map them without using rasters.

Next, we detail the data locations and formats for 5 subsets of data within the data folder. At any level of the folder organisation, data within a user-selected folder can be downloaded as a zip file.

1. Environmental rasters

Within the `/data/Environment` folder at the location above, rasters (~ 1GB total unzipped) are arranged in folders, one per region, and supplied as .tif files. In each region’s folder, a metadata file explains all known details for each variable. Within the `/data/Environment` folder, a README.txt file adds authors responsible for data preparation, and details of coordinate reference systems, units and raster cell sizes.

2. Presence-only data—locations and environmental samples

Within the `/data/Records/train_po` folder at the location above, there is one .csv file per region containing records for all species, and each is accompanied by a metadata file providing details for each column. Users will find that file formats are consistent across regions (Table 1).

3. Background data—locations and environmental samples

Ten thousand background (BG) samples are supplied for each region, as outlined in an earlier section. Within the `/data/Records/train_bg` folder at

[3] http://hdl.handle.net/1808/30582

[4] https://osf.io/kwc4v/

Table 1: Summary of data available across regions.

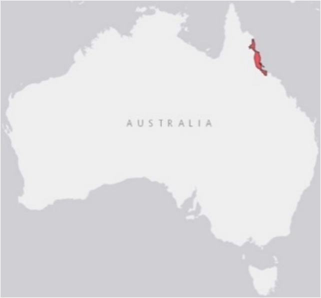
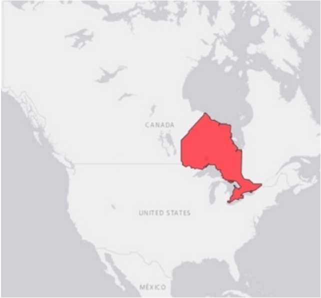
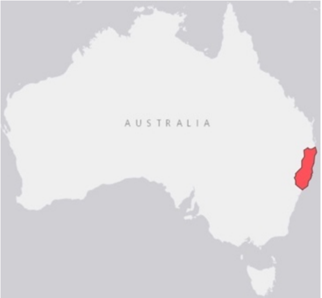
Code	Region details	Area ('000 km^2)	Area location – red polygons show locations within countries / continents	No. env vars (no. categorical)	Approx. grid cell resolution (m)	Biological groups & number species	Mean no. records per species		No.sites: PA
							PO	PA	
AWT	Australian Wet Tropics, Queensland, Australia	23.97		13 (0)	80	b: birds: 20	155	97	340
						p: vascular plants: 20	35	30	102
CAN	Ontario, Canada	979.34		11 (1)	1 000	birds: 30	253	1 282	14 571
NSW	North-east New South Wales, Australia	76.18		13 (1)	100	ba: bats: 7	27	76	570
						db: diurnal birds: 8	189	57	702
						nb: nocturnal birds: 2	134	142	1 137
						ot: open-forest trees: 8	42	164	2 075
						ou: open-forest understorey vascular plants: 8	21	358	1 309
						rt: rainforest trees: 7	9	212	1 036
						ru: rainforest understorey vascular plants: 6	18	93	909
						sr: small reptiles: 8	84	63	1 008

Table 1: Summary of data available across regions (continued).

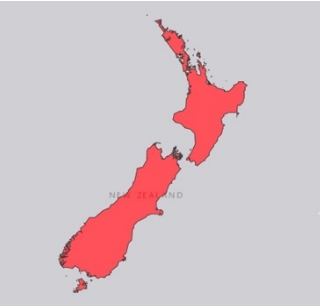
NZ	New Zealand	265.41		13 (2)	100	vascular plants: 52	59	1 801	19 120
SA	Continental Brazil, Ecuador, Colombia, Bolivia, and Peru, South America	12223.17		11 (0)	1 000	vascular plants: 30	74	12	152
SWI	Switzerland	39.56		13 (1)	100	trees: 30	1 170	810	10 013

Table 2: Information on sources of PO and PA data (initials refer to co-author names; all other acronyms defined elsewhere in text)

	PO data	PA data
AWT	Birds: supplied by SW. Incidental surveys. GPS locations therefore more accurate than many PO data. Plants: supplied by AF & KR. Herbarium data, cleaned and corrected by AF. Reliability codes for sites selected were 2 (15%), 3 (75%) and 4 (10%). This means that most records are accurate to within 3km and were collected within the 40 years prior to 2002.	Birds: supplied by SW. Field surveys. GPS locations. Plants: AF's survey sites. These have accurate locations (within 0.1km) and were collected in the 20 years prior to 2002.
CAN	Birds from the Ontario Nest Records database, Royal Ontario Museum (ROM). Supplied by M. Peck to FH. Temporal Span 1870-2002 (usually 1960-2001). Coordinates derived from map by ROM; some locations ground-truthed with GPS.	From Breeding Bird Atlas (BBA) for Ontario, provided by M. Cadman to FH.
NSW	8 biological groups supplied by SF. Fauna data from the Atlas of NSW Wildlife (a database of incidental sighting records); flora data: specimen records from both the University of New England Herbarium and the Sydney Herbarium (Royal Botanic Gardens). No information on collection dates or accuracy.	Supplied by SF. From designed surveys described elsewhere (Ferrier and Watson 1996; Pearce et al. 2001)
NZ	Plants, mostly trees and shrubs from indigenous forests. Supplied by JO, SW. Records from Allan Herbarium, managed by Manaaki Whenua -- Landcare Research	Supplied by JO, SW. Records from National Vegetation Survey databank (Wiser et al. 2001), nvs.landcareresearch.co.nz.
SA	Plant species from the family Bignoniaceae. Supplied by BL and LL. From the Missouri Botanical Garden database management system TROPICOS (http://www.mobot.org) and Lucia Lohmann (lohmann@mobot.org). Species localities were calculated by TROPICOS and by L. Lohmann using The Getty Thesaurus of Geographical Names Browser (http://shiva.pub.getty.edu).	Supplied by BL and LL. Survey data collected by Al Gentry over 22 years (1971-1993).
SWI	30 tree species Supplied by NEZ & TW From a forest vegetation data base containing 14 800 irregularly and non-systematically sampled forest vegetation relevés throughout Switzerland. Records start in 1904 and ends in 1995. The majority (95%) of the plots was collected after 1940, and ~60% of the data were sampled between 1960 and 1995. The individual authors had their own local sampling design or used preferential sampling techniques (details in Wohlgemuth 2012). Species cover estimation prevails as performance measure (98%) and follows the Braun-Blanquet approach (Braun-Blanquet 1964). The data is part of the European vegetation archive EVA (Chytrý et al. 2016). Around 14,100 relevés were selected from the original data base, targeting minimal data standards such as coordinates, and species extracted from these.	30 tree species Supplied by NEZ & MA Data extracted from the Swiss National Forest Inventory (NFI). PA data is collected on accessible sample plots at a regular 1 km point lattice across Switzerland. The data originate from the first national inventory, collected 1983-1985. For details, see Brassel and Lischke (2001) and EAFV (1988).

the location above, there is one .csv file per region, with columns exactly matching those for the PO data. Hence, PO and BG files can easily be combined for modeling, as needed. One README.txt file points to relevant files explaining the data for all regions and explains additional details specific to the background data setup.

4. Presence-absence data

PA data are intended for evaluation and supplied in two sets of files both at the location above: one (`/data/Records/test_env/`) containing the sampled environments for predicting to, and one (`/data/Records/test_pa/`) with the species data, in identical site (row) order to the environmental data. This two-file format reflects our original use of the data, keeping the evaluation data "blind" to the modeling (see comments on usage in the final section). Where data span more than one biological group (AWT, NSW), there are multiple files, one for each group.

In each of the two folders of .csv files, there is a README.txt file that points to relevant files explaining the data for all regions and providing details particular to the evaluation data.

5. Polygons of region extents

Polygons defining the extent (i.e. the borders) of each region are provided at the location above, in the `/data/Borders/` folder.

B. R package

Datasets 2 to 5, above, are also available in an R package, "disdat."[5] In the future we intend to submit it to CRAN.[6] The R package contains the data, func-

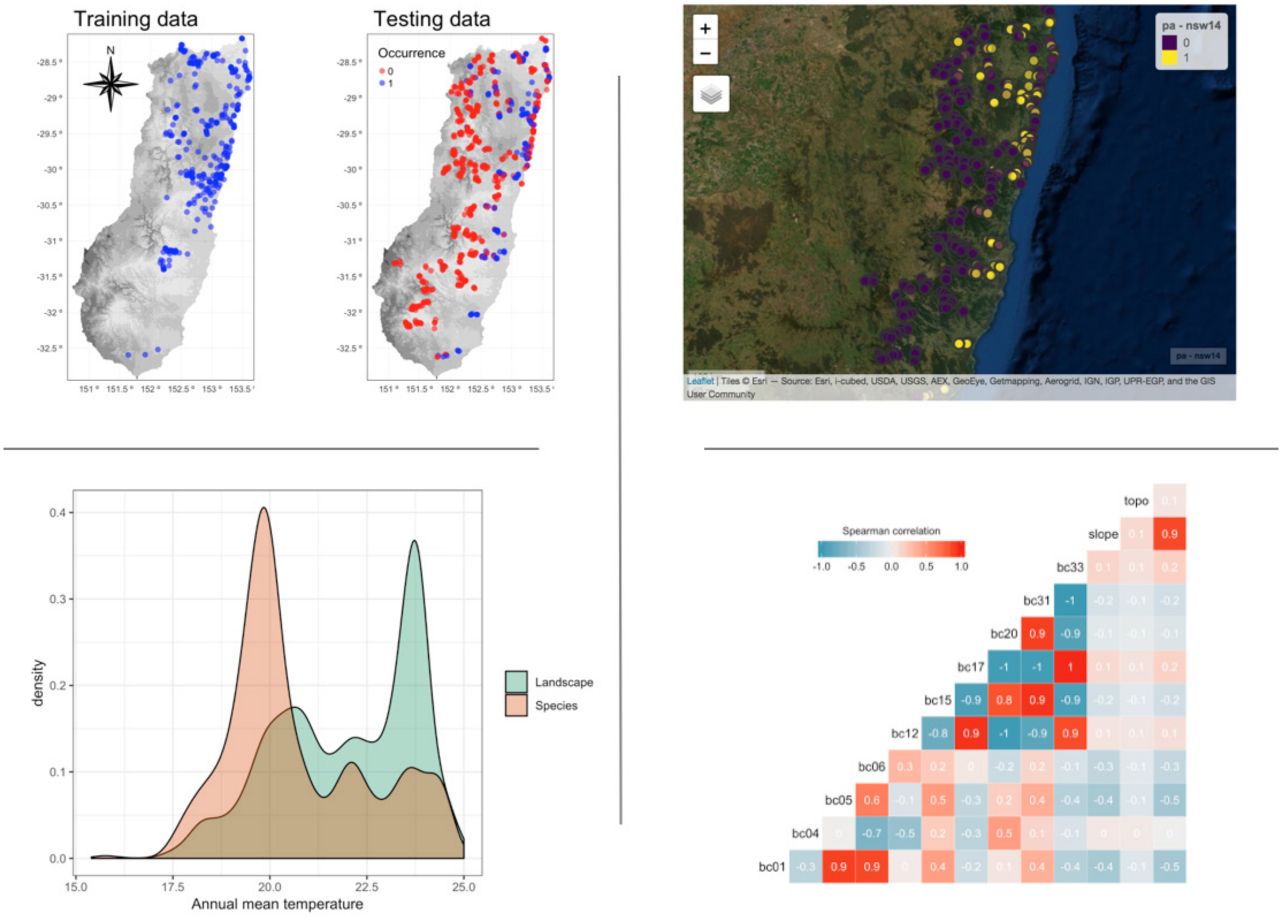


Figure 2: Examples of plots that can be achieved with functions supplied in the R package vignette. Top left: maps of species data, top right: an interactive map with PA site locations; Bottom left: density plot showing the distribution of PO data along one environmental gradient, compared with that of random points from the region; Bottom right: pairwise correlations between variable.

[5] https://github.com/rspatial/disdat/blob/master/README.md.
[6] https://cran.r-project.org/

tions to call the data, help files describing the data and pointing to this paper and its metadata, and two vignettes to assist modelers in data use.

Discussion of the Data and its Usage

Before we used these data for modeling, extensive efforts were made to prepare the data in an appropriate format. In retrospect—and as a comment for future data preparation exercises—some steps might have been done differently. For example, instead of removing duplicate records we could have marked records as "selected" and others as, for example, "not selected (duplicate in raster cell)". In addition, we reduced the species records to one per cell across regions with varying cell sizes. Instead across all regions a minimum distance between sites could have been set and applied to all. However, our cleaning is a common approach and not essentially flawed, so the data are still highly valuable. We no longer hold the original data, and our intention is to publish the dataset as used by publications listed in Supplementary Information 1, allowing comparison with these previous influential modeling efforts.

We note that the environmental data in the species files were extracted from rasters that we are also releasing. This extraction was done almost two decades ago. When we now, with current software, extract environmental data at those same site locations we find some very minor discrepancies in some datasets. These will likely have negligible effects on models. Nevertheless, we supply the values in the species files 'as is', because these are the data used for modeling in the most well-cited output of the NCEAS working group (Elith et al. 2006).

Much can be learned both about different methods and about properties of data by iterative modeling and evaluation. Readers can find analyses conducted to date on these data, in the publications summarized in Supplementary Information 1. One example of what we learned addresses bias in the species data. In 2002 there were very few published explorations of datasets like these, and their biases. This contrasts with the many published explorations and methods available now for handling bias in presence-only data (Kadmon et al. 2004; Phillips et al. 2009; Kramer-Schadt et al. 2013; Syfert et al. 2013; Warton et al. 2013; Bird et al. 2014; Boria et al. 2014; Dorazio 2014; Fithian et al. 2015; Stolar and Nielsen 2015; Qiao et al. 2017). The NCEAS group first explored whether the cleaned data could reasonably be used to predict species occurrence, without any bias treatment. The NCEAS modeling group demonstrated (Elith et al. 2006) that in some cases predictions had reasonable to very good accuracy, but that some regions' datasets were clearly hampered by bias. This led to subsequent work, particularly that of Phillips and co-authors (Phillips et al. 2009) who explored the extent and impact of bias in these data, and presented and tested the "target-group" approach for dealing with bias. In other words, our understanding of the problem of bias developed from our first iteration of modeling and our analyses of the outputs. We look forward to future insights gained from working with these data.

In the R package and associated vignettes, we present methods for exploring the supplied data to give insight into their properties. In the R package we provide a function for mapping all species, producing a "map book" of all PO and PA data for all regions. In the data visualisation vignette we provide code for mapping any given species (PO, PA data on a static map, plus an interactive map linked to satellite image data). In that vignette we also provide functions for exploring the distribution of sites in geographic and environmental space, and for analysing pairwise correlations between variables. Figure 2 illustrates some of the outputs that can be produced with these functions.

Our data are available on on OSF and GitHub as detailed in earlier, and are easy to download. We kindly request that each user (even students within teaching exercises) download the data or R package individually because some data providers would like to track data downloads, to enable reporting on data usage as required by their funding agencies.

Part of the value of this dataset is that independent PA data are available for evaluation of models fitted with PO data. In the publications shown in the table in Supplementary Information 1, the PA evaluation (test) data were kept independent as a "blind evaluation" set, that is, they were not used to tune models. In line with this setup, we have supplied the data in two distinct sets: 1) the environmental conditions at each evaluation site, which enables predictions to be made to the sites; 2) the actual PA observations at the evaluation sites for evaluation after modeling is complete. To facilitate future comparative research, we encourage users to provide clear documentation if they choose to use a different setup—for instance, tuning their models on some or all of the evaluation data.

Whilst we provide 10,000 background points for each region, recent research has shown that in some regions larger background samples may be required to sufficiently represent all environments (Renner et al. 2015). We supply the background data used in the main output of the NCEAS working group (Elith et al. 2006), though note that such datasets are easy to re-create by sampling the environmental raster data that we supply. They also may be sampled according to designs other than spatially at random. The use of random background data is justified if the PO occurrence data are a representative sample of the species' distribution (Phillips et al. 2009; Elith et al. 2011), a condition that is likely not satisfied for many species and regions. One alternative approach is to create and use a target group background sample (TGB; Phillips et al. 2009) where the target group contains all species in the identified group (e.g. all birds), including the species to be modeled. Depending on how the PO data are treated, the TGB sample might also need to be reduced to one sample per unique location. The "Modeling NCEAS data" vignette in the R package includes example code for making a TGB sample.

To replicate models in the main output of the NCEAS working group (Elith et al. 2006), users should read both the full paper and the associated appendix (freely available[7]). Since we were at that stage aiming to reflect expert use of models, different authors implemented different methods, starting with the data supplied here. As can be seen in the details recorded in that manuscript and its appendix, some modelers chose to use all predictors (with or without automated variable selection) whereas others chose a subset of the predictors based on pairwise correlations. Specific settings for each method are also documented in the appendix of Elith et al. 2006. Whilst code is not available to reproduce those models, in future work we plan to reproduce several of the models and provide fully documented R code for those algorithms, for reproducibility. In the meantime, and in order to help less experienced modelers, we also provide example code in the R package vignette "Modeling NCEAS data" for using the data for modeling and evaluation, applying one method across all species.

Acknowledgments

The working group "Testing alternative methodologies for modeling species' ecological niches and predicting geographic distributions" (project ID: 4980) was central to this project, funded and hosted by the National Centre for Ecological Analysis and Synthesis, Santa Barbara, California.

Competing Interests

The authors have declared that no competing interests exist.

References

Amano, T., and W. J. Sutherland. 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. Proc. R. Soc. B Biol. Sci. 280.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3:327–338.

Bird, T. J., A. E. Bates, J. S. Lefcheck, N. A. Hill, R. J. Thomson, G. J. Edgar, R. D. Stuart-Smith, S. Wotherspoon, M. Krkosek, J. F. Stuart-Smith, G. T. Pecl, N. Barrett, and S. Frusher. 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173:144–154.

Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275:73–77.

Brassel, P., and H. Lischke. 2001. Swiss National Forest Inventory: Methods and Models of the Second Assessment. Swiss Federal Institute WSL, Birmensdorf.

Braun-Blanquet, J. 1964. Pflanzensoziologie. Grundzüge der Vegetationskunde. Springer-Verlag Wein, New York.

Chytrý, M., S. M. Hennekens, B. Jiménez-Alfaro, I. Knollová, J. Dengler, F. Jansen, F. Landucci, J. H. J. Schaminée, S. Aćić, E. Agrillo, D. Ambarlı, P. Angelini, I. Apostolova, F. Attorre, C. Berg, E. Bergmeier, I. Biurrun, Z. Botta-Dukát, H. Brisse, J. A. Campos, L. Carlón, A. Čarni, L. Casella, J. Csiky, R. Ćušterevska, Z. D. Stevanović, J. Danihelka, E. D. Bie, P. de Ruffray, M. D. Sanctis, W. B. Dickoré, P. Dimopoulos, D. Dubyna, T. Dziuba, R. Ejrnæs, N. Ermakov, J. Ewald, G. Fanelli, F. Fernández-González, Ú. FitzPatrick, X. Font, I. García-Mijangos, R. G. Gavilán, V. Golub, R. Guarino, R. Haveman, A. Indreica, D. I. Gürsoy, U. Jandt, J. A. M. Janssen, M. Jiroušek, Z. Kącki, A. Kavgacı, M. Kleikamp, V. Kolomiychuk, M. K. Ćuk, D. Krstonošić, A. Kuzemko, J. Lenoir, T. Lysenko, C. Marcenò, V. Martynenko, D. Michalcová, J. E. Moeslund, V. Onyshchenko, H. Pedashenko, A. Pérez-

[7] http://www.ecography.org/sites/ecography.org/files/appendix/e4596.pdf

Haase, T. Peterka, V. Prokhorov, V. Rašomavičius, M. P. Rodríguez-Rojo, J. S. Rodwell, T. Rogova, E. Ruprecht, S. Rūsiņa, G. Seidler, J. Šibík, U. Šilc, Ž. Škvorc, D. Sopotlieva, Z. Stančić, J.-C. Svenning, G. Swacha, I. Tsiripidis, P. D. Turtureanu, E. Uğurlu, D. Uogintas, M. Valachovič, Y. Vashenyak, K. Vassilev, R. Venanzoni, R. Virtanen, L. Weekes, W. Willner, T. Wohlgemuth, and S. Yamalov. 2016. European Vegetation Archive (EVA): an integrated database of European vegetation plots. Appl. Veg. Sci. 19:173–180.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob. Ecol. Biogeogr. 12:1472–1484.

Dudik, M., and S. J. Phillips. 2009. Generative and discriminative learning with unknown labeling bias. Pp. 401–408 *in* Advances in Neural Information Processing Systems 21 (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.). Neural Information Processing Systems 2008, Neural Information Processing Systems Foundation, Inc.

EAFV. 1988. Schweizerisches Landesforstinventar: Ergebnisse der Erstaufnahme 1982-1986.

El-Gabbas, A., and C. F. Dormann. 2018. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41: 1161–1172.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, S. J. Phillips, K. S. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17:43–57.

Ferrier, S., and G. Watson. 1996. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Consultancy report prepared by the NSW National Parks and Wildlife Service for Department of Environment, Sport and Territories, Environment Australia, Canberra.

Fithian, W., J. Elith, T. Hastie, and D. Keith. 2015. Bias Correction in Species Distribution Models: Pooling Survey and Collection Data for Multiple Species. Methods Ecol. Evol. 6:424–438.

Garzon-Lopez, C. X., L. Bastin, G. M. Foody, and D. Rocchini. 2016. A virtual species set for robust and reproducible species distribution modelling tests. Data Brief 7:476–479.

Graham, C. H., J. Elith, R. J. Hijmans, A. Guisan, A. T. Peterson, and B. A. Loiselle. 2008. The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45:239–247.

Guisan, A., C. H. Graham, J. Elith, F. Huettmann, and . NCEAS Species Distribution Modelling Group. 2007. Sensitivity of predictive species distribution models to change in grain size: insights from a multi-models experiment across five continents. Divers. Distrib. 13:332–340.

Harrell, F. E. 2001. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis. Springer Verlag, New York.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction, second edition. 2nd ed. Springer-Verlag, New York.

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93:679–688.

Hortal, J., A. Jiménez-Valverde, J. F. Gómez, J. M. Lobo, and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847–858.

Isaac, N. J. B., and M. J. O. Pocock. 2015. Bias and information in biological records. Biol. J. Linn. Soc. 115:522–531.

Kadmon, R., O. Farber, and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14:401–413.

Kissling, W. D., J. A. Ahumada, A. Bowser, M. Fernandez, N. Fernández, E. A. García, R. P. Guralnick, N. J. B. Isaac, S. Kelling, W. Los, L. McRae, J.-B. Mihoub, M. Obst, M. Santamaria, A. K. Skidmore, K. J. Williams, D. Agosti, D. Amariles, C. Arvanitidis, L. Bastin, F. De Leo, W. Egloff, J. Elith, D. Hobern, D. Martin, H. M. Pereira, G. Pesole, J. Peterseil, H. Saarenmaa, D. Schigel, D. S. Schmeller, N. Segata, E. Turak, P. F. Uhlir, B. Wee, and A. R. Hardisty. 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93:600–625.

Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, B. Schröder, J. Lindenborn, V. Reinfelder, M. Stillfried, I. Heckmann, A. K. Scharf, D. M. Augeri, S. M. Cheyne, A. J. Hearn, J. Ross, D. W. Macdonald, J. Mathai, J. Eaton, A. J. Marshall, G. Semiadi, R. Rustam, H. Bernard,

R. Alfred, H. Samejima, J. W. Duckworth, C. Breitenmoser-Wuersten, J. L. Belant, H. Hofer, and A. Wilting. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19:1366–1379.

Lawson, C. R., J. A. Hodgson, R. J. Wilson, and S. A. Richards. 2014. Prevalence, thresholds and the performance of presence–absence models. Methods Ecol. Evol. 5:54–64.

Munafò, M. R., B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1:0021.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2009. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. The National Academies Press, Washington, DC.

Pearce, J. L., K. Cherry, M. Drielsma, S. Ferrier, and G. Whish. 2001. Incorporating expert knowledge and fine-scale vegetation mapping into statistical modelling of faunal distribution. J. Appl. Ecol. 38:412–424.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19:181–197.

Qiao, H., A. T. Peterson, L. Ji, and J. Hu. 2017. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. Methods Ecol. Evol. 8:1804–1812.

Reddy, S., and L. M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. J. Biogeogr. 30:1719–1727.

Renner, I. W., J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. 2015. Point process models for presence-only analysis. Methods Ecol. Evol. 6:366–379.

Roberts, D. W., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929.

Stolar, J., and S. E. Nielsen. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Divers. Distrib. 21:595–608.

Syfert, M. M., M. J. Smith, and D. A. Coomes. 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PloS One 8:e55158.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. PloS One 8:e79168.

Warton, D. I., and L. C. Shepherd. 2010. Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. Ann. Appl. Stat. 1383–1402.

Wiser, S. K., P. J. Bellingham, and L. E. Burrows. 2001. Managing biodiversity information: development of New Zealand’s National Vegetation Survey databank. N. Z. J. Ecol. 25:1–14.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan, and NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14:763–773.

Wohlgemuth, T. 2012. Swiss forest vegetation database. Pp. 340–340 *in* Dengler J., J. Oldeland, F. Jansen, M. Chytry, J. Ewald, M. Fickh, F. Glöckler, G. Glopez-Gonzalez, R. K. Peet, and J. H. J. Schaminée. (editors) Vegetation databases for the 21st century.

Zuckerberg, B., F. Huettman, and J. Friar. 2010. Chapter 3: Proper Data Management as a Scientific Foundation for Reliable Species Distribution Modeling. Pp. 45–70 *in* Predictive species and habitat modeling in landscape ecology. Springer New York.