

RESEARCH ARTICLE

Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models

Petra Fühding-Potschkat¹  | Holger Kreft¹ | Stefanie M. Ickert-Bond²

¹Biodiversity, Macroecology and Conservation Biogeography, Faculty of Forest Sciences, University of Göttingen, Göttingen, Germany

²Department of Biology and Wildlife & UA Museum of the North, University of Alaska Fairbanks, Fairbanks, Alaska, USA

Correspondence

Petra Fühding-Potschkat, Biodiversity, Macroecology and Conservation Biogeography, Faculty of Forest Sciences, University of Göttingen, Büsgenweg 1, 37077 Göttingen, Germany.
Email: fuehrding@gmail.com

Funding information

Georg-August-Universität Göttingen

Abstract

Digital point-occurrence records from the Global Biodiversity Information Facility (GBIF) and other data providers enable a wide range of research in macroecology and biogeography. However, data errors may hamper immediate use. Manual data cleaning is time-consuming and often unfeasible, given that the databases may contain thousands or millions of records. Automated data cleaning pipelines are therefore of high importance. Taking North American *Ephedra* as a model, we examined how different data cleaning pipelines (using, e.g., the GBIF web application, and four different R packages) affect downstream species distribution models (SDMs). We also assessed how data differed from expert data. From 13,889 North American *Ephedra* observations in GBIF, the pipelines removed 31.7% to 62.7% false positives, invalid coordinates, and duplicates, leading to datasets between 9484 (GBIF application) and 5196 records (manual-guided filtering). The expert data consisted of 704 records, comparable to data from field studies. Although differences in the absolute numbers of records were relatively large, species richness models based on stacked SDMs (S-SDM) from pipeline and expert data were strongly correlated (mean Pearson's r across the pipelines: .9986, vs. the expert data: .9173). Our results suggest that all R package-based pipelines reliably identified invalid coordinates. In contrast, the GBIF-filtered data still contained both spatial and taxonomic errors. Major drawbacks emerge from the fact that no pipeline fully discovered misidentified specimens without the assistance of taxonomic expert knowledge. We conclude that application-filtered GBIF data will still need additional review to achieve higher spatial data quality. Achieving high-quality taxonomic data will require extra effort, probably by thoroughly analyzing the data for misidentified taxa, supported by experts.

KEYWORDS

automated data cleaning pipelines, data quality, expert data, GBIF, species distribution modeling

TAXONOMY CLASSIFICATION

Ecoinformatics

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Digitally accessible species records from global data-sharing networks like the Global Biodiversity Information Facility (GBIF) provide the basis to address a wide range of biodiversity-related questions in ecology, biogeography, and other disciplines (e.g., Guralnick et al., 2007; Meyer et al., 2016; Soberón & Peterson, 2004). Such databases and data-sharing networks represent a valuable source of knowledge in which individual researchers and institutions worldwide invested considerable amount of time and resources (Baskauf et al., 2016; Guralnick et al., 2018; Wiczorek et al., 2012). However, since the circumstances and standards under which these records were collected and digitized are usually unknown, a user must assess whether the data quality provided meets the requirements of the research question (Beck et al., 2013; Sterner & Franz, 2017). Consequently, this demands data cleaning tools (hereafter: DC tool) to standardize data and identify and remove data errors. Thus, developing appropriate DC tools is a long-standing goal of biodiversity informatics (e.g., Araújo & Guisan, 2006; Chapman et al., 2000; Kadmon et al., 2004).

Data errors occur mainly along three dimensions: taxonomy, space, and time (Meyer et al., 2016). They may significantly affect common downstream analyses such as the accuracy of species distribution models (SDMs, e.g., Gueta & Carmel, 2016; Tassarolo et al., 2017; Hijmans & Elith, 2019; Zizka et al., 2019). In the taxonomic dimension, resolving misspellings (Zermoglio et al., 2016) and reconciling the synonymy of taxonomic names (Alroy, 2002; Wortley & Scotland, 2004) pose a significant challenge. The related widespread and particularly challenging problem is misidentified specimens, estimated at 50% for tropical plant specimens (Goodwin et al., 2015) and ranging from 5% to nearly 60% in the Zoological Record database (Meier & Dikow, 2004). In the spatial dimension, errors in and low precision of coordinates, for example, from rounding of the decimal digits, swapped latitude and longitude, missing coordinates, or coordinates with zero-values are common data quality problems (e.g., Otegui et al., 2013; Töpel et al., 2017; Yesson et al., 2007). Lower geospatial accuracy is frequently assumed for older records than for those collected more recently (Tassarolo et al., 2017; Zizka et al., 2020). Stropp et al. (2016) showed, for instance, that conspicuous records of flowering plants collected in Africa before the 1960s were filtered out due to poor data quality. Another issue associated with older records is that the probability increases that populations no longer exist at a given sampling location over time due to natural or anthropogenic reasons (Meyer et al., 2016).

Even for experts, identifying and resolving data quality issues manually is in many cases unfeasible, given that datasets typically contain thousands to millions of records. Therefore, selective DC strategies based on well-explained instructions and automated DC tools that reproducibly generate high-quality data are especially in high demand for inexperienced users (Zizka et al., 2019). Downstream applications such as conventional SDMs depend on these data quality (e.g., Araújo et al., 2019; Guisan et al., 2017; Raes

& Aguirre-Gutiérrez, 2018). Data scientists and biodiversity informaticians approached the development of DC solutions from several angles: (1) DC tools that generally solve thematically limited requirements, like retrieving, evaluating, formatting, completing, and organizing data. This type of DC solution was implemented in the widely used *Tidyverse* "umbrella" package (Wickham et al., 2019). The solution was also included in specialized packages such as *CoordinateCleaner* (Zizka et al., 2019), *rgbif* (Chamberlain, 2020), and the GBIF web application (GBIF.org, 2020). (2) Manuals supporting the preparation of data for SDMs. Particular *R* packages are an integral part of such manuals (e.g., Chapman, 2005; Guisan et al., 2017; Hijmans & Elith, 2019). The manuals consist of verbal explanations and coded instructions, which the user can apply (e.g., per package *dismo*, Hijmans & Elith, 2020). While the newly developed and recently updated methods for automated cleaning of records are promising, their effect on commonly applied SDMs remains poorly examined (see Hijmans et al., 2017; Schmidt-Leubuh et al., 2013; Zizka et al., 2020).

Pipelines play an important role in the scientific domain when, for example, biodiversity data from different sources such as herbarium vouchers and observations need to be combined for analysis. In this study, we investigated the performance of six pipelines (P1 to P6) using various DC tools and how these pipelines affected downstream SDMs. We used North American *Ephedra* species as the model organisms (Ephedraceae, Gnetales; Cutler, 1939; Stevenson, 1993, Figure 2, A to C; Table S1) and GBIF as the data source. With over 2.1 billion species records worldwide, GBIF is the largest and one of the most frequented public providers of biodiversity data. It is often the primary data source for many researchers (Guralnick et al., 2018; Hobern et al., 2019; Zizka et al., 2020). Thus, we selected the GBIF records as input to the pipelines. In this context, we address three questions:

1. How do the pipelines differ in their performance? We expect that different DC tools will generate different result datasets.
2. How do differences in pipeline data affect downstream diversity models and maps (observed, predicted)? We expect the pipeline datasets to differ in the resulting models (single species and stacked SDMs, hereafter: S-SDM) and maps.
3. How does the pipeline data—after being cleaned by the pipelines—differ from the expert data (observed and predicted), assuming that the expert data represent the most accurate *Ephedra* environmental and geographical range? We expect the quality of the pipeline data to differ from the expert data. The differences will be measurable (occurrences and correlations) in the models and maps.

We analyzed to which extent the data from the different pipelines led to different species constellations and numbers in the grid cells and visualized the differences in diversity maps created from S-SDMs. Finally, we discuss how realistic the results from GBIF data and expert data reflect the environmental or geographical extent of the *Ephedra* species' ranges.

2 | MATERIALS AND METHODS

In North America, *Ephedra* species are characteristic components of arid and semi-arid regions of the southwestern USA and Mexico (Hollander & VanderWall, 2009; Loera et al., 2015). They occur from the Death Valley to about 2500m in the Rocky Mountains (Stevenson, 1993). The species share a morphologically reduced, uniform growth habit with mostly leafless, photosynthetic stems (Ickert-Bond & Renner, 2016). Specimens are collected frequently, as shown by the record numbers of the public providers (e.g., GBIF: 46,384 records worldwide), and high-quality expert data are available for the New World species (Ickert-Bond, 2003). The coordinates served as the proxy for the *Ephedra* species' characteristic locations (response variables), from which we developed species SDMs and genus S-SDMs for North America.

We monitored changes in similarities and correlations using the validated records from P1 to P6 and the expert data (observed occurrences, hereafter: L1; Table 2). From L1, we developed L2 and L3 data of the North American *Ephedra* species and their occupied grid cells (per pipeline and the expert data). L2 included the grid cell numbers

an *Ephedra* species occupied, and L3 counted the concurrent *Ephedra* species per grid cell. L4 data comprised the correlations of the observed occupied grid cells. The L5 data (pipeline and expert) included the predicted distribution in S-SDMs across the pipelines and expert data (L2/L4, and L5: Spatial autocorrelation by Moran's I and correlation between two random variables by Pearson's r) (Figure 3).

2.1 | Data pipelines

Ensuring comparability across six pipelines, the process chain of filters provided identical conditions to optimize the provider data (See Table 1, the filters of the pipelines). The chain consisted of (1) selecting and retrieving data from GBIF, (2) standardizing the records by filtering, and (3) correcting or removing data errors (Figure 1, Table 2). At each pipeline step, we employed one or more DC tools (e.g., Chapman, 2005; Hijmans & Elith, 2019; Zizka, 2019). The selected tools (e.g., GBIF web application, written instructions, or R packages) or their most recent updates were released between 2005 and 2020 and are free of charge. In some pipelines, the three

TABLE 1 Pipeline filter summary for standardization and error removal

Categories	Filter	Requirement	Rationale
STD	Country range	Spatial	North America: Mexico and the USA
STD	Intraspecific rank	Taxonomic	Required rank: species (Claridge et al., 1997; Reydon & Kunz, 2019), infraspecific ranks (e.g., subspecies, hybrids) to be omitted.
STD	Collection years	Temporal	1945 to 2020, as older records are more likely to contain erroneous coordinates (Zizka et al., 2020).
STD	Basis of record	Consistency	Specimens and observations.
STD	Occurrence status	Consistency	Presence data.
FPS	Non-North America-native <i>Ephedra</i> species	Taxon	All non-native <i>Ephedra</i> species that are allocated to the North American countries either by mistake or are artificially introduced, for example, to botanical gardens.
FPS/REC	Zero or missing coordinates	Spatial	Zeros and missing values may represent records with data entry errors. Missing values will cause error messages in <i>ade4</i> .
REC	Longitude and latitude are equal	Spatial	Equal longitude and latitude may represent records with data entry errors.
DUP	Duplicate records	Consistency	Duplicate records that may represent, for example, record copy errors.
FPS	Country capitals	Spatial	Records that may contain the coordinates of the country capital.
FPS	Country centroids	Spatial	Records that may contain the centroid coordinates of the country.
FPS	GBIF headquarters	Spatial	Records that may contain the coordinates of the GBIF headquarters.
FPS	Biodiversity institutions	Spatial	Records that may contain the coordinates of biodiversity institutions where the herbarium voucher is stored.
FPS	Geographic outliers	Spatial	Geographic outliers that may represent misidentified specimens.
REC	Urban areas	Spatial	Records from urban areas that may represent old data or vague locality descriptions.
REC	dd.mm to dd.dd conversion errors	Spatial	Records with ddmm to dd.dd conversion error (misinterpretation of the degree sign as decimal delimiter).
REC	Rasterized collections	Spatial	Records with a significant proportion of coordinates that might have a low precision.
FPS	"Manual" removal of false positives	Consistency	False positives that have been overlooked by automated error removal, based on the knowledge that they are in the records.

Note: Categories: DUP, duplicate records; FPS, false positives; REC, recording errors; STD, standardization.

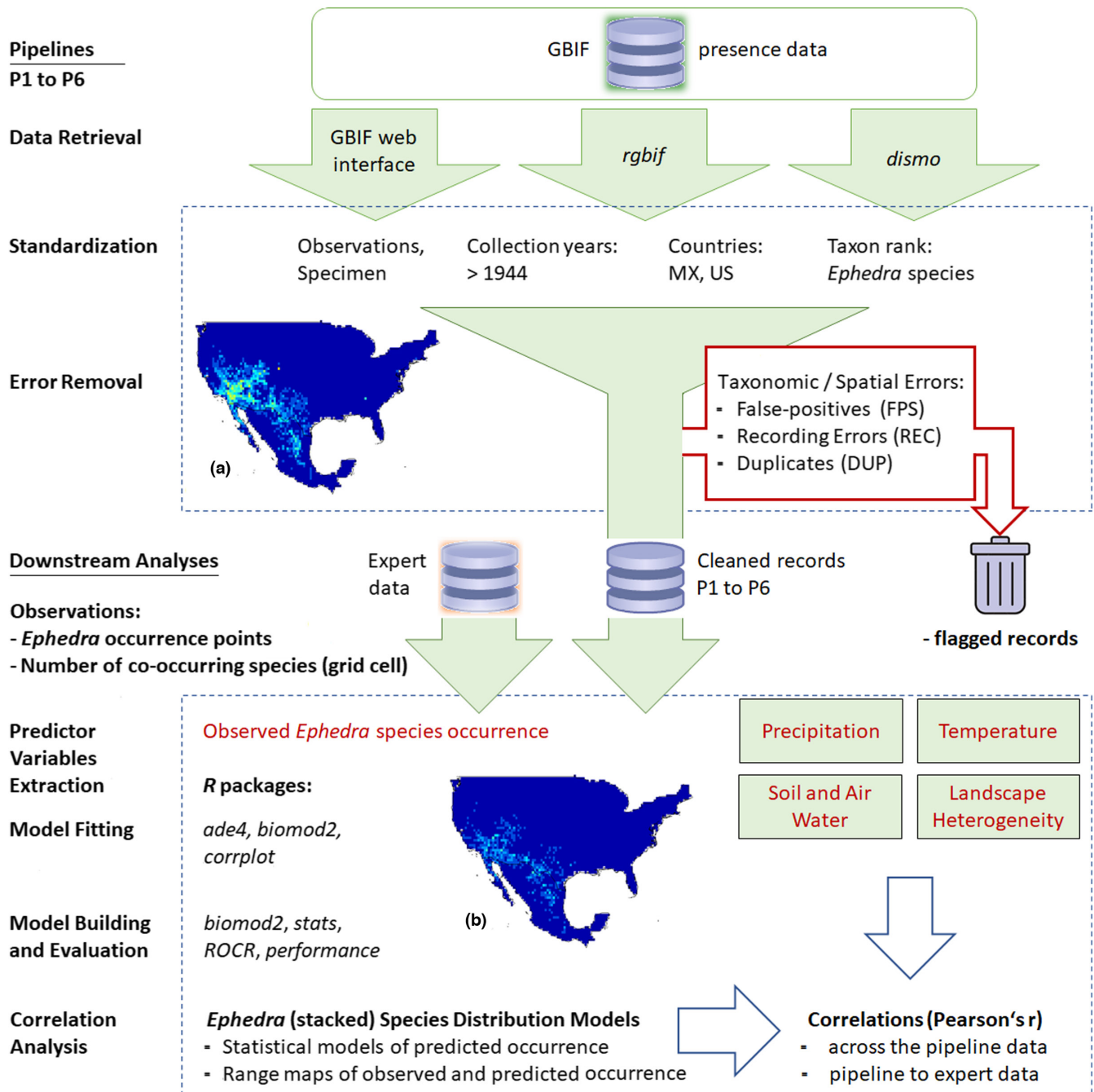


FIGURE 1 Workflow of the pipelines and the downstream analyses. The pipelines' part comprised the following sections: Data Retrieval, Standardization, and Error Removal. The Downstream Analysis featured the Predictor Variables Extraction, the Model Fitting, the Model Building (SDMs, S-SDMs) and Evaluation, and the Correlation Analysis developed from the pipeline data P1 to P6 and the expert data. *R* packages used in the course of the workflow are in italics. (a) Observed species distribution from GBIF P1 data. (b) Observed species distribution from expert data. Filter categories: DUP, Duplicate records; FPS, False positives; REC, Recording Errors.

steps were performed by one ("three-in-one") DC tool. In the setup of the process chain, we followed the data cleaning recommendations given by the respective DC tool's authors and pertinent best-practice guidelines (Araújo et al., 2019; Guisan et al., 2017).

We retrieved data from GBIF (gbif.org, 2020) on November 18, 2020, in four different ways: (1) The filter "*Ephedra* L." (hereafter: GBIF (I)) retrieved 46,384 records for P5, P6, and the P0 benchmark data using the "three-in-one" GBIF web application (GBIF, 2020a). (2)

The filter set "*Ephedra* L. specimens of North America, from 1945 to 2019" (hereafter: GBIF (II)) selected 9484 records for the P1 process chain using the web application (GBIF, 2020b). In both cases, the data were downloaded with the web application. (3) *rgbif*, a "three-in-one" tool, employed its integrated functionality to standardize the P2 and P3 data and retrieved 6687 GBIF records into the userspace. (4) *dismo* selected 46,384 GBIF records for P4 and retrieved them into the userspace. (Details see Table 2).

TABLE 2 Results of the pipelines' data cleaning performance, compared to the P0 benchmark dataset (summary table)

Pipeline datasets	P1	P2	P3	P4	P5	P6	P0 benchmark
Input: Data retrieved by	GBIF (II)	<i>rgbif</i>	<i>rgbif</i>	<i>dismo</i>	GBIF (I)	GBIF (I)	GBIF (I)
Number of records retrieved	9484	6687	6687	46,384	46,384	46,384	46,384
Non-native <i>Ephedra</i> species outside North America	NA	NA	NA	32,495, rem	32,495, rem	32,495, rem	32,495, flg
Number of records passed to the standardization	NA	NA	6687	13,889	13,889	13,889	13,889
Data standardized by	GBIF	<i>rgbif</i>	<i>rgbif</i> , <i>dplyr</i>	<i>R</i> code	GBIF (I), <i>dplyr</i>	GBIF (I), <i>dplyr</i> , <i>CC</i> , <i>R</i> code	
North America-sampled <i>Ephedra</i> specimens (MX, US)	9484	6687	6687	13,889	13,889	13,889	13,889
Occurrence status: presence	default	default	default	default	default	default	default
Non-native <i>Ephedra</i> specimens in North America	31, ret	0	0	55, rem	55, rem	55, rem	55, flg
Not identifiable specimens in North America (e.g., genus level, fossil)	296, ret	0	0	501, rem	501, rem	501, rem	501, flg
North America-native, taxon rank: species	9010	6687	6678	13,240	13,240	13,240	13,240
Intraspecific ranks	147, ret	0	0	704 rem	704 rem	704 rem	704 flg
Collection years: >1944	9484	6687	6687	9560	9560	9560	9560
Collection years: < 1945	NA	NA	NA	4329 rem	4329 rem	4329 rem	4329 flg
Basis of record: observations, specimens	9484	6560	6560	13,762	13,762	13,762	13,762
Other basis of records	NA	127, ret	127, rem	NA	127, rem	127, rem	127 flg
Number of records passed to the data cleaning	NA	NA	6,560	8,300	8,173	8,173	
Data cleaned by	NA	<i>rgbif</i>	<i>rgbif</i> , <i>CC</i>	<i>R</i> code	<i>dplyr</i> , <i>R</i> code	<i>CC</i> , <i>R</i> code	
NULL coordinates (Missing values)	2592, ret	rem	rem	1852, rem	1758, rem	1766, rem	5978, flg
Zero coordinates	8, ret	rem	rem	8, rem	8, rem	8, rem	8, flg
Longitude and latitude are equal	8, ret	8, ret	8, rem	12, rem	12, rem	12, rem	22, flg
Duplicate records (species, longitude, latitude, year, month, day)	1086, ret	1226, ret	1182, rem	1,31, rem	998, rem	1000, rem	3584, flg
Country capitals	1, ret	NA	NA	1, ret	1, ret	1, rem	1, flg
Country centroids	9, ret	8, ret	23, rem	23, ret	23, ret	23, rem	23, flg
GBIF headquarters	NA	NA	NA	NA	NA	NA	NA
Biodiversity institutions	33, ret	19, ret	19, rem	36, ret	36, ret	36, rem	36, flg
Geographic outliers	12, ret	12, ret	12, rem	35, ret	35, ret	35, rem	35, flg
Sea coordinates	146, ret	67, ret	67, rem	228, ret	228, ret	61, rem	228, flg
Urban areas	193, ret	165, ret	165, ret	298, ret	298, ret	2, ret	298, flg
dd.mm to dd.dd conversion errors	202, ret	202, ret	0	278, ret	278, ret	0, rem	278, flg
Rasterized collections, possibly reduced coordinate precision	56, ret	56, ret	56, rem	56, ret	56, ret	41, rem	56, flg
Unidentified false positives (manually identified and removed)	2, ret	2, ret	2, ret	1, rem	2, rem	2, rem	2, flg
Number of records passed to the data finalization	NA	NA	5,198	5,396	5,395	5,196	
Data standardized and finalized by	NA	NA	<i>R</i> code	<i>R</i> code	<i>R</i> code	<i>R</i> code	
Native <i>Ephedra</i> species, sample size < 50 occ points	53, ret	9, ret	9, rem	9, rem	9, rem	9, rem	93, flg
Output: Final number of cleaned records	9484	6687	5189	5387	5386	5187	13,889

Color code key	Data retrieval	Data cleaning
<i>dismo</i>	Y	NA
basic <i>R</i> code	NA	Y
<i>coordinateCleaner</i>	NA	Y
<i>dplyr</i>	NA	Y
GBIF (I)	Y	Y
GBIF (II)	Y	Y
<i>rgbif</i>	Y	Y

Note: The color-coded cells of P1 to P6 datasets indicate the activity of a particular DC tool (color code see below). The blue cells of the P0 benchmark indicate the number of *Ephedra* records in GBIF, quantified by standardization and error category. Records which did not comply with the standardization conditions or were erroneous in the context of this study were flagged (flg). Since several standardization conditions and errors coincided in the same record, the number of removed records did not correspond to the sum of the identified errors. The P1, P2, and P3 data retrieval tools partially standardized the data and eliminated several errors ("three-in-one" tools). Thus, the number of records retrieved differed significantly from P4 to P6, and P0. The removed records in these pipelines could only be reconstructed as differences of subcategories (e.g., in-scope countries, collection year, null and zero coordinates) in comparison to P0. The difference between P3 and P2 resulted from the added *dplyr* and *CC* packages, which increased standardization and removed still more erroneous records. Using the added packages ensured more insight into data cleaning. Abbreviation: CC (→ P3/P6) = *R* package *CoordinateCleaner*.

We created the P0 data for comparison. It served as the benchmark of standardization and errors, delivered by the GBIF data, which the DC tools could have removed in the pipelines. However, P0 was not itself a pipeline nor was it part of any pipeline. We performed an inventory of the dataset and the data errors that might influence the quality of the downstream models (Table 2, P0 column). Using P0, we could identify questionable records and the degree of feasibility to which each pipeline removed such records. After data

retrieval, further data cleaning was performed in P3, P4, P5, and P6 by basic *R* code (R Core Team, 2013), the *dplyr* package (of Tidyverse, Wickham et al., 2019), and the *CoordinateCleaner* (Zizka, 2019; Zizka et al., 2019), in different combinations (Table 2). We selected records of taxon rank "species" (Claridge et al., 1997; Reydon & Kunz, 2019), filtered for North America (Mexico, USA), and collection years 1945 to 2020 (Zizka et al., 2020). As the basis of records, we selected specimens and observations. During error removal, we focused on

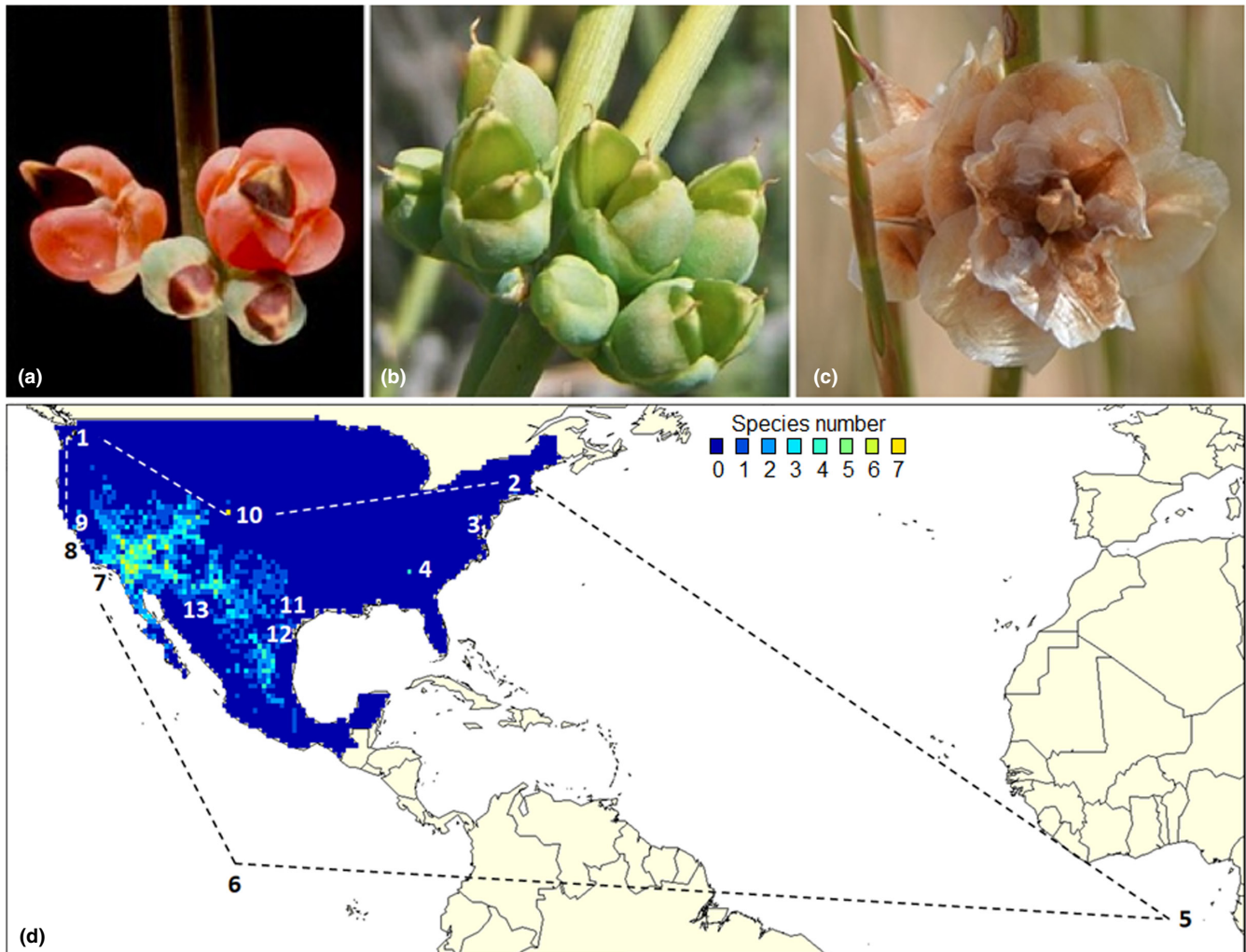


FIGURE 2 (a–c) North America-native *Ephedra* specimens (female specimens with seeds). *Ephedra antisiphilitica*, *E. nevadensis*, and *E. trifurca* (left to right). (d) Examples of taxonomic and spatial errors identified in the *Ephedra* data. Filter categories of the following markers: False positives. Markers 1, 8, and 9 were specimens from shops in Seattle and Berkeley. Markers 3, 4, 10, and 11 were non-native species from botanical gardens and scientific institutes. Marker 2 pointed to a North America-native species at the University of Connecticut, NY. Markers 5 to 7 showed coordinate errors that the verbatim locality description can only identify. The species at markers 12 and 13 were misidentified, as the documented species do not occur naturally at these localities. The data for the map derived from the P1, post-cleaning (L3, number of co-occurring species). Color coding of the map: P1 observed distribution (see Figure 4).

taxonomic and spatial errors (Meyer et al., 2016), such as non-native specimens, missing or zero values, and sea coordinates. We also removed false-positive records reporting, for example, occurrences at biodiversity institutions, and geographic outliers. From the P0 evaluation, we were aware of two false-positive occurrences (Figure 2, Marker 2) hidden in the data. We found these errors challenging to be recognized by any tool. Therefore, we removed one of these errors in P4, and two in P5 and P6, using basic R code. As coordinates with three or fewer decimal places often indicate they were obtained from grid maps (Zizka et al., 2019), we permitted only validated coordinates with no less than four decimal places. However, this precision was not required for the modeling. The *CoordinateCleaner* identified specimens of urban areas and flagged them for scrutiny. We searched for duplicates based on the variables: species, coordinates, and collection date, respectively, and removed them. Finalizing the

process chains, we excluded native species for which the sample size was lower than 50 occurrences to avoid biased models and maps (Guisan et al., 2017; Hijmans & Elith, 2019). (Usage of the tools in the pipelines, see Table 2). At the end of the pipelines, we examined the retained records and errors in the pipelines' datasets in comparison to P0 (data at L1).

2.2 | Downstream analysis

Data from examination of physical herbarium specimens and field studies (Ickert-Bond, 2003) represented the most realistic environmental and geographical range ("gold standard", Araújo et al., 2019) of the genus *Ephedra* in North America. The expert dataset comprised 4081 records of New World *Ephedra* specimens from herbaria

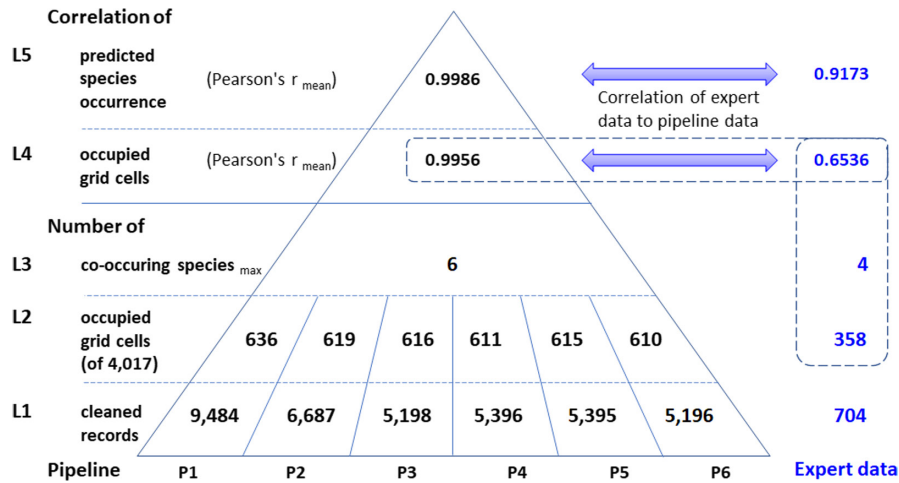


FIGURE 3 Information condensing pyramid of the pipelines and the expert data (L1 to L5: Condensing levels of the data). The data show an increasingly higher correlation from the bottom to the top of the pyramid, which results from data transformations into an increasingly higher condensed species occurrence information state. The 704 expert data occurrences (L1) were allocated into 358 grid cells (L2, with a maximum of four co-occurring species, L3). The correlation of 0.6536 (L4, mean Pearson's r of pairings [P1 to P6/expert]) was compared to the mean of the pairings P1 to P6. At this level (L4), the minimum Pearson's r -value of the occupied grid cells from pipeline data was .9920 (pair: P1/P6), and the maximum Pearson's r value was .9999 (pair: P4/P5). At the L5 level, the minimum Pearson's r value was .9951 (pair: P1/P6), and the maximum Pearson's r value was 1.0000 (pair: P4/P5). Dashed box: Expert data comparison numbers, L2 to L4.

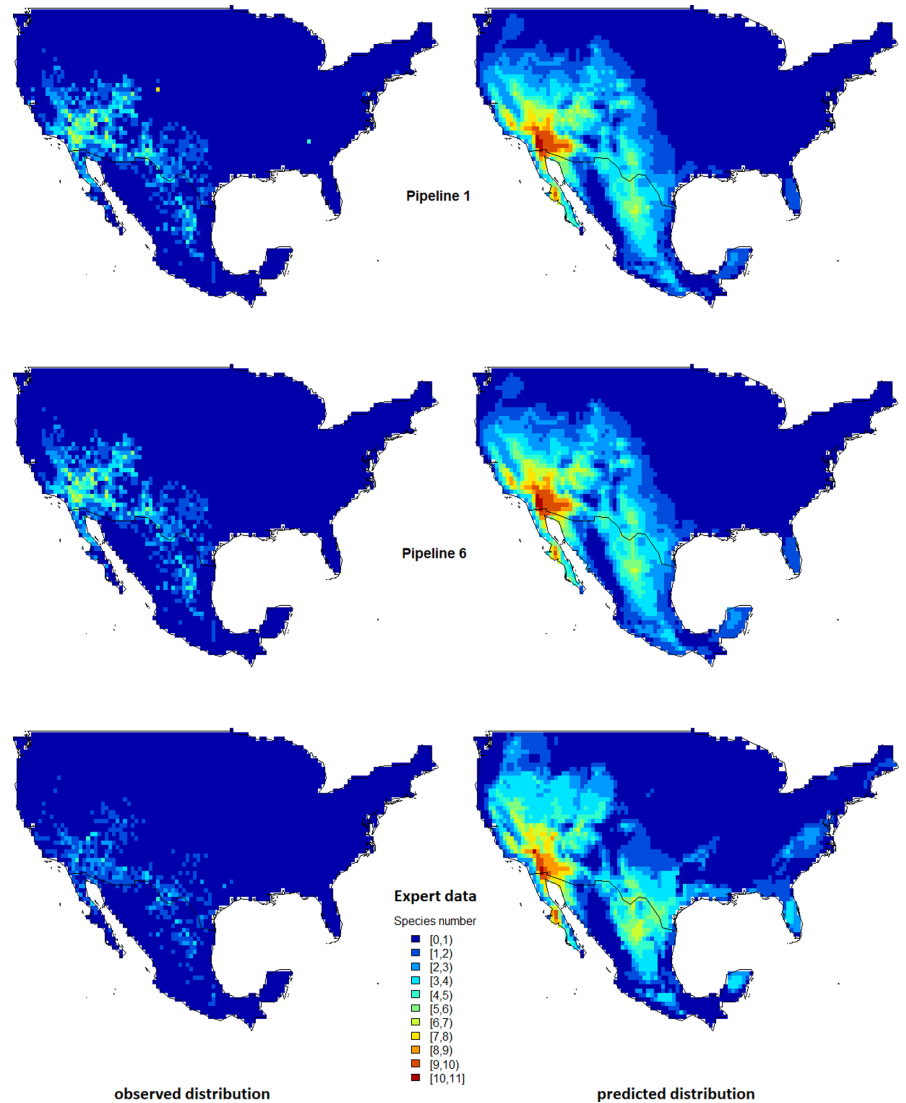


FIGURE 4 Stacked species distribution maps based on cleaned GBIF data from pipelines P1, P6, and expert data. Depicted are the maps of the least cleaning P1 and the most cleaning P6 that show only minor differences (the maps from the other pipeline data are close to P6). The control data map from the expert data shows differences to the pipelines. Left: Observed distribution (L2 data). Point-occurrences after passing the pipelines, allocated to grid cells of a stacked range map of all *Ephedra* species. The expert map shows less occupied grid cells ($n = 358$) than P1 ($n = 636$) resulting in a smaller range. Right: Map of the predicted probability of species from S-SDMs (L5 data). The color keys show highly correlated patterns of each data quality (P1, P6, and expert data: 0 to 12 species, Pearson's $r = .9173$).

with large holdings of *Ephedra* in both North and South America (e.g., ARIZ, ASU, HUH, NY, RM, SGO, SI, TEX, UC, UNAM, US; herbarium acronyms according to Thiers, 2022). A total of 704 records of 12 *Ephedra* species (L1) were selected for North America; however, they were not processed in a pipeline. We applied standardization conditions only for comparability. The records contained confirmed taxa, examined coordinates, and detailed locality descriptions comparable to field-collected data. We considered an overlap of 90 records of 13,889 from GBIF and the expert dataset negligible. As *Ephedra* is adapted to dry environments, we imported 19 temperature and precipitation variables from the CHELSA climatology (Karger et al., 2017), elevation data as a proxy for landscape heterogeneity (GMTED, 2020), and plant-available water data (Zhang et al., 2018). From their habitat description (e.g., Cutler, 1939; Stevenson, 1993), we assumed the selected environmental data being ecologically relevant.

For the SDMs and S-SDMs, we created a grid of 4017 cells across Mexico and the USA (30 arc minutes, WGS84) using `wrld_simple` (R package `maptools`, Bivand et al., 2022) and `raster` (Hijmans et al., 2017). The grid size reasonably showed the co-occurring species, which was not the case on different scales. We aggregated the environmental data to the grid resolution (`sp` package, version 1.4-5, Bivand et al., 2013; Pebesma & Bivand, 2005) and extracted the values for each occurrence (`raster`; Hijmans & van Etten, 2021). We built a presence-absence table, creating a random selection of pseudo-absences for each *Ephedra* species using the R package `biomod2` (Thuiller et al., 2016). We tested the localities where *Ephedra* species were not recorded (R package `ecospat`, Di Cola et al., 2017). We anticipated environmental conditions to cause absence (Loera et al., 2015; Stevenson, 1993), making sure that the localities used for fitting the model represented the requirements of the species across North America (Training area, Guisan et al., 2017). We summed up the species present in the grid cells as the number of co-occurring species. (L2, L3).

We identified the contributing predictors (using R packages `ade4`, Bougeard & Dray, 2018 and `corrplot`, Wei et al., 2017). From the 21 variables, we selected a subset of reasonably uncorrelated variables per species using `biomod2` (Table S2; Guisan et al., 2017; Thuiller et al., 2016). Reasonably uncorrelated refers to being below the recommended threshold of 0.7 (Dormann et al., 2013). As goodness-of-fit evidence, we used the Akaike Information Criterion (AIC; Johnson & Omland, 2004), and Tjur's R^2 (Coefficient of Discrimination for binary outcomes; R package `performance`, Lüdecke et al., 2021) to identify the variables with the highest impact (Table S2). Finally, we fitted logistic regression models for the *Ephedra* occurrences using `glm` as the model and "binomial" as the distribution family. The threshold value of a high-performance index (0.9, Guisan et al., 2017) was used to evaluate the predictive accuracy of the model, particularly the Receiver Operating Characteristic Curve (ROC) and the area under the curve (AUC) (R packages `biomod2`, Thuiller & Lafourcade, 2019, and `ROCR`, Sing et al., 2015). We stacked the predictions of the 12 *Ephedra* species resulting from the different pipelines as well as the expert data to S-SDMs (without

using thresholds; Biber et al., 2020; Calabrese et al., 2014; Guisan et al., 2017). The correlations between the observed and the predicted *Ephedra* occurrences informed how strongly the differences between the pipelines and the expert data affected the respective SDMs and S-SDMs (L5).

We inspected spatial autocorrelation (L2/L4: grid occupation, L5: predicted distributions) using the Moran's I coefficient (R package `spdep`, Bivand et al., 2015). We computed the correlations of the observed and predicted *Ephedra* occurrences in two pipelines (the least cleaned data, P1, and the most cleaned data, P6) and the expert data using Pearson's r (R package `rstatix`, Kassambara, 2020). Ultimately, we visualized them as map pairs (Figure 4); and to adequately represent the species richness in the maps, we chose 11 breaks (R package `classInt`, Bivand, 2022) for the maximum possible co-occurring species.

3 | RESULTS

The GBIF web interface using GBIF (I) filters and `dismo` retrieved 46,384 unstandardized and uncleaned, globally distributed *Ephedra* datasets. The GBIF web interface using GBIF (II) filters retrieved 9484 partially standardized *Ephedra* records from North America. `rgbif` retrieved 6687 somewhat standardized specimen records from North America and already removed significant spatial errors. (Download results see Table 2). The three tools stopped after the data retrieval.

3.1 | P0 benchmark data

About 13,889 P0 records represented the unstandardized and uncleaned GBIF North American *Ephedra* data. A total of 1979 specimens were collected or observed in Mexico (14.2%) and 11,910 in the USA (85.8%). The majority of species records consisted of North America-native *E. viridis* (19.0%), *E. aspera* (14.4%), *E. californica* (14.1%), *E. nevadensis* (13.3%), *E. trifurca* (11.9%), and *E. torreyana* (8.7%), a total of 81.4% for six species. Another six native species, *E. antisiphilitica* (4.4%), *E. funerea* (2.4%), *E. fasciculata* (1.8%), *E. pedunculata* (1.5%), *E. compacta* (1.3%), and *E. cutleri* (1.1%) totaled 12.5%. The remaining 6.1% were non-native (55 taxonomic false positives of South American and Eurasian origin) or indeterminate specimens (499 specimens of genus *Ephedra* L.). Several standardization conditions and errors coincided in the same record. Thus, the number of removed records did not correspond to the sum of the identified errors. About 5187 records (37.3%) were flagged as fit-for-use for the downstream analyses. Around 8702 records (63.7%) were marked for removal due to one or more significant errors. Missing coordinates (5978 records, 43.1%) represented the majority of identified data errors, followed by the sampling year (4329 records, 31.1%, were older than 1945) and the duplicate records (3584 records, 25.8%). About 220 records showed coordinates in bodies of water. With two exceptions, the non-native *Ephedra* species were, for example, found in botanical

gardens and scientific institutes (e.g., Atlanta Botanical Garden; Figure 2d, locality markers 3, 4, 10, and 11). As a few non-native species contain medicinally active substances, they were reported with two records from a shop in Berkeley (*E. sinica*, Figure 2d, locality markers 8 and 9) and one record from an herbal product shop in Seattle (*E. sinica*, Figure 2d, locality marker 1). We detected *E. nevadensis* at the University of Connecticut (Figure 2d, locality marker 2), yet this species is native to the Southwestern United States. Three records revealed misplaced taxa by comparing the verbatim locality description with the coordinates. These errors were not identified by a tool, only by scrutiny. Locality marker 12 referenced a misidentified specimen (*E. distachya*, Figure 2d) that does not naturally occur in Coahuila, Mexico. The specimen that locality marker 13 referenced (*E. trifurcata*, Figure 2d) might be a misspelling of *E. trifurca* (P0 results, see Table 2, Table S1).

3.2 | Expert data

Five hundred seventy-seven of 2251 specimens were collected or reviewed from Mexico (25.3%) and 1674 specimens from the US (75.4%). After standardization, 704 records remained (210 records of Mexican specimens, 494 records of US specimens). After standardization, the majority of records (65.2%) were allocated to *E. aspera* (22.3%), *E. trifurca* (21%), *E. fasciculata* (11.4%), and *E. antisiphilitica* (10.5%). The other eight species, *E. viridis* (7.1%), *E. californica* (5.4%), *E. torreyana* (5.1%), *E. funerea* (4.5%), *E. compacta* (3.7%), *E. pedunculata* (3.3%), *E. nevadensis* (2.3%), and *E. cutleri* (1.4%) totaled 32.8% of the standardized records. The remaining 14 records (2%) were of other taxonomic ranks.

3.3 | Effects of differences in the pipeline data on diversity models

P1 and P2 were partly standardized in their process chain. GBIF (II) of P1 met four out of five standardization requirements. Explicit error removal did not occur; however, P1 implicitly removed 3386 missing coordinate records as a side effect of the standardization. It left 2592 missing coordinates records, 296 indeterminate records, and 33 South American and Eurasian species in the P1 dataset. P2's *rgbif* met three standardization requirements, but the resulting data still contained infraspecific ranks. *rgbif* standardized the P2 data partly, using the parametrized standardization criteria, and, in addition, the built-in error exclusion parameter of invalid coordinates was employed. Except for excluding missing values in the coordinates, P2 removed no other spatial errors.

P3, P4, P5, and P6 continued their respective process chains. The pipelines removed between 43.1% and 45.3% of all spatial error types (e.g., the complete subset of 5986 missing coordinates records, see Table 2). P3 used the *dplyr* and *CoordinateCleaner*, providing 5189 records to the downstream analyses. In P4, we fully standardized the data, using instructions explained in a tutorial (Hijmans & Elith, 2019) and basic R code. P4 provided 5387 records to the downstream

analyses. In P5, we standardized the data and removed errors, using basic R code and the *dplyr*. P5 provided 5386 records to the downstream analyses. P6 used instructions from Chapman (2005) translated to basic R code and *dplyr* functionality to handle taxonomic errors. The *CoordinateCleaner* removed spatial errors. P6 identified 5187 fit-for-use records for the downstream analyses. Due to not meeting the sampling size criteria, we manually removed *Ephedra coryi* records from the pipelines. At the end of the pipelines, the records for the downstream analyses varied considerably and ranged from 9484 (P1) to 5187 (P6) (L1) (Table 2).

The cleaned datasets differed by 4288 (P1 vs. P6), and the number of occupied grid cells by 26 grid cells (maximum). We observed similarly clustered occupancy patterns in the distribution maps regardless of the pipeline since most records were allocated to the same grid cells per species. The occupied grid cells in the stacked *Ephedra* range maps varied between 636 and 610 (P1 vs. P6 data). Comparisons of highly correlated occupied grid cells (mean Pearson's *r* across the pipelines: .9956) were confirmed by highly correlated maps of observed *Ephedra* distribution with well-defined clusters (Figure 3, and Figure 4, P1 and P6 map pairs). Moran's *I* confirmed the spatially clustered patterns of the *Ephedra* species (observed P1/P6 Moran's *I*: 0.144, observed expert data's Moran's *I*: 0.087, *p*-value: significant) (L2/L4). *Ephedra californica* occurrences occupied identical grid cells across all six pipelines; therefore, the Pearson correlation coefficient was 1. For the other 11 *Ephedra* species, the occupancy of the grid cells varied slightly across the pipelines, depending on the respective pipelines compared. For example, in *E. fasciculata*, P1 differed from P6 with 49 versus 53 occupied grid cells (92.5% identical occupancy), while the occupancy in P2 and P3 in *E. antisiphilitica* was again identical (Pearson's *r* = 1). The evaluation of the S-SDMs showed that the grid cell occupancy patterns (observed occurrences) continued in the species distribution maps (predicted occurrences). Correlograms based on residual analysis are listed in Figure S2.

Post-pipelines, we found that the *ade4* indicated coordinates with missing values as invalid in records containing this error type, hence, may also be regarded as a testing point for missing values in the coordinates. (Note that we did not intervene in the data cleaning in P1 by GBIF (II). Thus, records with missing values in coordinates were preserved).

The final number of predictors for the species ranged from 4 (*Ephedra aspera*) to 10 (*Ephedra viridis*) (Table S2). The area under the curve (AUC) scored from 0.9355 (*Ephedra antisiphilitica*) to 0.9990 (*Ephedra nevadensis*) (AUC mean: 0.9825). The AIC decreased to a stable minimum value in the variable's combination tests, indicating the best possible model performance compared to the other variable combinations. Therefore, we considered our models as adequately accurate to describe the distribution of the *Ephedra* species with the identified explanatory variables. The differences in the pipelines had a minor effect on the correlations, models, and maps at L4 and L5. At level L4, the mean Pearson's *r* of the occupied grid cells across the pipelines was 0.9956 (P1/P6 pair: 0.9920, minimum; P4/P5 pair: 0.9999, maximum). The high correlation led to maps of observed

Ephedra distribution that showed also only insignificant differences (Figure 4, P1 and P6 observed distribution). Across the six pipelines, the predicted probability of occurrence from the S-SDMs indicated high correlations (mean Pearson's $r = .9986$, Figure 3, L5). Figure 4 displays the maps of the predicted distribution based on the S-SDMs.

3.4 | Differences between pipeline data and expert data

The 704 expert data occurrences (L1) were allocated into 358 grid cells, with a maximum of four co-occurring species (L3). Across the pipelines, 294.5 of the average 630.5 grid cells (46.7%) showed occupancy by one species, compared to 265 of 358 grid cells (74.0%) of the expert data. 42.6 of the grid cells showed occupancy by four species (6.7%), compared to the maximum of four species (1.1%) of the expert data. Ten grid cells showed occupancy by the maximum of six species (1.6%) in the pipeline data (L2).

The correlations differed clearly between the pipelines and the expert data. At level L4, the mean Pearson's r of the occupied grid cells for pipeline data correlated to the expert data was .6536 (L4: Figure 3). The correlation of the predicted occurrence probabilities in the S-SDMs showed a mean Pearson's r of .9173. Across the different pipelines and the expert data, the observed diversity in the maps from the S-SDMs showed a large *Ephedra* diversity center in Southern California that continued to the North into Arizona and Nevada, and to the South into the states of Baja California and Sonora, Mexico with a predicted *Ephedra* diversity greater than seven species. A second diversity center emerged across the state of Texas, USA, and continued into the states of Chihuahua, Coahuila, Nuevo León, and Tamaulipas, Mexico, with a predicted *Ephedra* diversity of up to seven species. (L5). The diversity patterns in the expert data, although similar in shape, were less distinct (Figure 4).

4 | DISCUSSION

We analyzed the data cleaning performance of six different pipelines for digital point-occurrence records and their effects on species distribution models, a common downstream application in macroecology. The six pipelines differed significantly in the number of accepted species, errors removed, and remaining records for analysis (Table 2, Table S1). For example, P6 removed the most significant number of records, approximately twice as many records as the least cleaning pipeline P1. Data from P1 differed from the other group by hosting 17 non-native species in addition to the 12 natives, all of which were removed by the other pipelines. P1 also retained false-positive coordinates (e.g., sea, country capitals and centroids, biodiversity institutes, herbal shops), geographic outliers, and duplicates, which were removed to different degrees by the pipelines of the other group (Table 2) (Question 1).

Due to the low complexity of the data cleaning environment, P1 and P2 required only little effort to get their pipelines installed.

Both pipelines did not achieve the standardization and error elimination anticipated to reduce unwanted effects in the downstream analyses. P1 identified potential shortcomings in the data only in a few cases due to the limited options of the GBIF filter application. In contrast, P3 to P6 were more demanding in the required know-how, mainly when using the *R* packages and preparing the respective user environments but offered a more substantial functionality (Table 2). The *R* packages performed the data cleaning well for coordinate errors that rendered records unusable for use in diversity models. Generalist packages like the *dplyr* and specialists like the *CoordinateCleaner*, especially in combination, reliably identified problematic records with missing values and false-positive occurrences such as biodiversity institutes or country centroids. Accurate distribution data are essential for any SDM and the many comparable downstream analyses (Araújo & Guisan, 2006; Chapman et al., 2000; Kadmon et al., 2004; Zizka et al., 2020). Therefore, the main aim of well-designed pipelines is to efficiently and automatically generate cleaned data tailored to the specific research question (Zizka et al., 2020; Table 1). We mainly focused on comparing the outcomes of different pipelines that used well-known data retrieval or DC tools to answer this question. The standardization filters served to unify the record structure across the pipelines. Although older herbarium vouchers or observations are as valuable as recent vouchers since they may document both a historical status and biodiversity changes over time (Meyer et al., 2016), the "collection year, older than 1945" filter, for example, was implemented to standardize the data but also to reduce expected general coordinate imprecisions up-front. However, removing taxonomic and spatial errors was at the core of the pipeline data for the model fitting and model building and the respective tools.

4.1 | Influence of different data cleaning solutions on downstream analyses

Removing the non-native species, which consisted of only a few specimens, reduced the number of cleaned records only slightly (per species and overall). The non-native *Ephedra* species had no noticeable effect in the occupied grid cells as co-occurring species. They were concentrated in a few places and in small numbers of species only (P1, Figures 3 and 4: observed distribution). The low level of differences was confirmed by reasonably high correlation coefficients, which continued to even higher correlation coefficients regarding the predicted probability of species in S-SDMs (L1 to L5: Figure 3). Removing the missing value records in the pipelines was essential for the downstream analyses. The model fitting tool issued error messages when identifying any in the provided data (*ade4*). Although we included the duplicate records filter in determining the number of duplicate records in the data, duplicate records did not affect the fitted models (Question 2).

The tested pipelines offer automated data cleaning in a standardized and reproducible manner. Pipeline P1 supports all users but produces data that still contain serious taxonomic and spatial errors.

In contrast, the pipelines P2 to P6, which help users with some programming experience (Zizka et al., 2019, 2020), produce data qualities where many errors were eliminated and which seem suitable for diversity model use (SDMs and S-SDMs).

4.2 | Significant differences of the expert data and the GBIF data

The P1 data differed noticeably from the expert data, for example, in the species composition (P1 data: 29 species vs. expert data, and P2 to P6 data: 12 species), the number of records per species, the number of occupied grid cells after the observations were allocated to gridded range maps (Figure 3, L2), and the number of co-occurring species. P2 to P6 differed less from the expert data. (Question 3). The aim of collating data for SDMs is to avoid bias and inaccuracies in taxonomic and distribution data, and an effective means of overcoming bias and inaccuracies is to build data from field studies (Araújo et al., 2019; Chapman, 2005). Well-maintained expert data support both the aims and provide an alternative to field studies. A less maintained data alternative, biodiversity records from GBIF, are free of charge but with limitations in data quality due to several known and unknown errors. Expert and GBIF data form the data layer (Bakshi, 2012; Vetter, 1990). However, the critical difference between expert data and GBIF data is that the expert data may be used unprocessed as input to the data modeling workflow as there are no data errors to be expected. For the GBIF data, an additional data cleaning process chain needs to be included in the workflow so that the data modeling can be meaningfully linked to the data layer. Consequently, a user of GBIF data always has to plan for an additional effort for the data cleaning design, which includes the functional structure of the target data that is fit for use, and a pipeline to obtain it (Wirth & Hipp, 2000; Zizka et al., 2019).

4.3 | A major issue: misidentified specimens that still hide in the dataset

Comparing the quantities of the GBIF pipelines' analysis data and the expert data shows that the expert data are roughly 11.8% or about one-eighth of the GBIF data (mean). From this ratio, we may assume that there are still many errors in the pipeline data, hence, the visible differences in the maps (Figure 4). This point opens the question of how realistic the GBIF data is. No pipeline detected taxonomic issues such as misidentifications or false positives like non-native specimens in the data due to a lack of information about their distributional status. For differently determined specimens of the same origin, given to other institutes and handled in isolation from their parent specimens, Nicolson (2019) provided a technical solution. We used expert know-how to assess the likeliness of taxonomic identities in recorded localities as there presently is no tool that possesses this functionality (Figure S1). Developing a tool that resolves this issue might be challenging considering the many names,

from synonyms to misspellings (Zermoglio et al., 2016). A correction method that was already introduced is that a data owner directly changes false positives identified in individual cases by notifying the provider. Generally, with the present interfaces to GBIF, it cannot be avoided that misidentified taxa enter into the databases by, for example, citizen scientists. Interfaces that prevent taxonomic or spatial errors before entering a public provider must be designed.

5 | CONCLUSION

Our results suggest that the P1 data show more differences from P2 to P6 data than within this group. Depending on the pipeline, one-third (P1) to two-thirds (P6) of the GBIF records were classified as unsuitable for biodiversity analyses. Importantly, differences in the pipeline data did not translate into significant differences in downstream SDMs and S-SDMs, suggesting remarkable robustness of these analyses toward data cleaning differences. The increasingly condensed information from the occurrence data led to ever stronger correlations across the pipelines. Three aspects emerged from the study. First, data from the GBIF web application require further cleaning. Second, the R packages reliably removed incorrect or dubious coordinates. Therefore, choosing the right DC tools depends on the researcher's skills. Third, it is challenging to identify misidentified specimens in the public data providers. To overcome this difficulty, we suggest new processes to identify misidentified specimens or prevent new misidentified specimens from being entered into the public data providers. Consequently, programmers developing new data cleaning packages should consider the requirements for data cleaning, notably as the *CoordinateCleaner* eliminates most spatial errors.

AUTHOR CONTRIBUTIONS

Petra Führding-Potschkat: Conceptualization (lead); data curation (lead); formal analysis (lead); funding acquisition (lead); investigation (lead); methodology (lead); project administration (lead); resources (equal); software (equal); supervision (lead); validation (equal); visualization (lead); writing – original draft (lead). **Holger Kraft:** Conceptualization (supporting); data curation (supporting); formal analysis (supporting); funding acquisition (supporting); investigation (supporting); methodology (supporting); project administration (supporting); resources (equal); software (equal); supervision (supporting); validation (equal); visualization (supporting); writing – original draft (supporting). **Stefanie M. Ickert-Bond:** Conceptualization (supporting); data curation (supporting); formal analysis (supporting); funding acquisition (supporting); investigation (supporting); methodology (supporting); project administration (supporting); resources (equal); supervision (supporting); validation (equal); visualization (supporting); writing – original draft (supporting).

ACKNOWLEDGMENT

We thank Pedro Tarroso and an anonymous reviewer for their helpful suggestions and comments on the earlier versions of the manuscript.

We also acknowledge statistical advice of Patrick Weigelt and fruitful discussion with the members of the Biodiversity, Macroecology, and Biogeography group. Open Access funding enabled and organized by Projekt DEAL.


CONFLICT OF INTEREST

The authors involved in the preparation of this manuscript have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

P0 benchmark data and pipelines P4 (*dismo*-retrieved), P5, and P6 data: 46,384 worldwide distributed *Ephedra* records from GBIF, <https://doi.org/10.15468/dl.2eg5ab>. (GBIF, 2020a). Pipeline P1 data, filtered by the GBIF web application: 9484 North America-distributed *Ephedra* records from GBIF, <https://doi.org/10.15468/dl.r2cg62>. (GBIF, 2020b). Pipelines P2 and P3 data (*rgbif*-retrieved): 6687 North America-distributed *Ephedra* records from GBIF, https://datadryad.org/stash/share/QspgKk8RRIEXK6grxNgdzde8KmfVOJc4_N6cfly_bQ. North American *Ephedra* Expert data, 704 North America-distributed *Ephedra* records, <https://datadryad.org/stash/share/7X7EDIZlIgLjkyFOXJoqYEvOER9k3q8vDic2CZN2jE>.

ORCID

Petra Fühding-Potschkat  <https://orcid.org/0000-0003-2838-9874>

REFERENCES

- Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences of the United States of America*, 99, 3706–3711. <https://doi.org/10.1073/pnas.062691099>
- Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5, eaat4858. <https://doi.org/10.1126/sciadv.aat4858>
- Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33, 1677–1688. <https://doi.org/10.1111/j.1365-2699.2006.01584.x>
- Bakshi, K. (2012). Considerations for big data: Architecture and approach. In *2012 IEEE aerospace conference*. <https://ieeexplore.ieee.org/abstract/document/6187357>
- Baskauf, S. J., Wiczorek, J., Deck, J., & Webb, C. O. (2016). Lessons learned from adapting the Darwin Core vocabulary standard for use in RDF. *Semantic Web*, 7, 617–627. <https://doi.org/10.3233/SW-150199>
- Beck, J., Ballesteros-Mejia, L., Nagel, P., & Kitching, I. J. (2013). Online solutions and the 'Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions*, 19, 1043–1050. <https://doi.org/10.1111/ddi.12083>
- Biber, M. F., Voskamp, A., Niamir, A., Hickler, T., & Hof, C. (2020). A comparison of macroecological and stacked species distribution models to predict future global terrestrial vertebrate richness. *Journal of Biogeography*, 47, 114–129. <https://doi.org/10.1111/jbi.13696>
- Bivand, R. (2022). *classInt*: Choose Univariate Class Intervals. <https://r-spatial.github.io/classInt/>, <https://github.com/r-spatial/classInt/>
- Bivand, R. S., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., & Blanchet, G. (2015). *Package spdep*. The Comprehensive R Archive Network. <https://www.yumpu.com/en/document/view/9283478/package-spdep-the-comprehensive-r-archive-network>
- Bivand, R., Lewin-Koh, N., Pebesma, E., Archer, E., Baddeley, A., Bearman, N., Bibiko, H. J., Brey, S., Callahan, J., Carrillo, G., Dray, S., Forrest, D., Friendly, M., Giraudoux, P., Golicher, D., Rubio, V. G., Hausmann, P., Hufthammer, K. O., Jagger, T., ... Johnson, K. (2022). *R package 'maptools'*. <https://r-forge.r-project.org/projects/maptools/>. Accessed May 25, 2022.
- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). Springer. <https://asdar-book.org/>
- Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in R with the *ade4* package. *Journal of Statistical Software*, 86, 1–17. <https://doi.org/10.18637/jss.v086.i01>
- Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, 23, 99–112. <https://doi.org/10.1111/geb.12102>
- Chamberlain, S. (2020). *rgbif*: Interface to the global biodiversity information facility API ver. 3.2.0. <https://docs.ropensci.org/rgbif/>
- Chapman, A. D. (2005). *Principles and methods of data cleaning – primary species and species-occurrence data, version 1.0*. Report for the Global Biodiversity Information Facility, Copenhagen.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* (Vol. 9, p. 13). SPSS Inc.
- Claridge, M. F., Dawah, H. A., & Wilson, M. R. (1997). *Species: The units of biodiversity*. Chapman and Hall Ltd.
- Cutler, H. C. (1939). Monograph of the North American species of the genus *Ephedra*. *Annals of the Missouri Botanical Garden*, 26, 373–428. <https://doi.org/10.2307/2394299>
- Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R. G., Hordijk, W., Salamin, N., & Guisan, A. (2017). *ecospat*: An R package to support spatial analyses and modeling of species niches and distributions. *Ecography*, 40, 774–787. <https://doi.org/10.1111/ecog.02671>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36, 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- GBIF. (2020a). *GBIF occurrence download*. <https://doi.org/10.15468/dl.2eg5ab>
- GBIF. (2020b). *GBIF occurrence download*. <https://doi.org/10.15468/dl.r2cg62>
- GBIF.org. (2020). *GBIF home page*. <https://www.gbif.org>
- GMTED. (2020). *GMTED digital elevation data, elevation above sea level (mn30_grd.zip)*. https://topotools.cr.usgs.gov/gmted_viewer/viewer.htm
- Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. I., & Scotland, R. W. (2015). Widespread mistaken identity in tropical plant collections. *Current Biology*, 25, R1066–R1067. <https://doi.org/10.1016/j.cub.2015.10.002>
- Gueta, T., & Carmel, Y. (2016). Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics*, 34, 139–145. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge University Press.
- Guralnick, R. P., Hill, A. W., & Lane, M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, 10, 663–672. <https://doi.org/10.1111/j.1461-0248.2007.01063.x>
- Guralnick, R., Walls, R., & Jetz, W. (2018). Humboldt Core—toward a standardized capture of biological inventories for biodiversity

- monitoring, modeling and assessment. *Ecography*, 41, 713–725. <https://doi.org/10.1111/ecog.02942>
- Hijmans, R. J., & Elith, J. (2019). *Spatial distribution models*. <https://www.rspatial.org/sdm/SDM.pdf>
- Hijmans, R. J., & Elith, J. (2020). *dismo: Species distribution modeling*. <https://rspatial.org/raster/sdm/>
- Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). Package 'dismo'. *Circles*, 9, 1–68. <https://cran.microsoft.com/snapshot/2018-04-14/web/packages/dismo/dismo.pdf>
- Hijmans, R.J., & van Etten, J. (2021). *raster: Geographic data analysis and modeling*. R package version 3.4-10. <http://CRAN.R-project.org/package=raster>
- Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A., van Huis, E., Kim, E.-S., McGeoch, M., Naicker, I., Navarro, L., Noesgaard, D., Price, M., Rodrigues, A., Schigel, D., Sheffield, C. A., & Wiczorek, J. (2019). Connecting data and expertise: A new alliance for biodiversity knowledge. *Biodiversity Data Journal*, 7, e33679.
- Hollander, J. L., & VanderWall, S. B. (2009). Dispersal syndromes in North American *Ephedra*. *International Journal of Plant Sciences*, 170, 323–330. <https://doi.org/10.1086/596334>
- Ickert-Bond, S. M. (2003). *Systematics of New World Ephedra L. (Ephedraceae): Integrating of morphological and molecular data* [Ph.D. dissertation, Tempe, Arizona State University].
- Ickert-Bond, S. M., & Renner, S. S. (2016). The Gnetales: Recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. *Journal of Systematics and Evolution*, 54, 1–16. <https://doi.org/10.1111/jse.12190>
- Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*, 19, 101–108. <https://doi.org/10.1016/j.tree.2003.10.013>
- Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14, 401–413.
- Karger, D. N., Conrad, O., Böhrner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific data*, 4, 1–20. <https://doi.org/10.1038/sdata.2017.122>
- Kassambara, A. (2020). *rstatix: Pipe-friendly framework for basic statistical tests*. R package version 0.6.0. <https://rpkgs.datanovia.com/rstatix/>
- Loera, I., Ickert-Bond, S. M., & Sosa, V. (2015). Ecological consequences of contrasting dispersal syndromes in New World *Ephedra*: Higher rates of niche evolution related to dispersal ability. *Ecography*, 38, 1187–1199. <https://doi.org/10.1111/ecog.01264>
- Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). *performance: An R package for assessment, comparison and testing of statistical models*. *Journal of Open Source Software*, 6, 3139. <https://doi.org/10.21105/joss.03139>
- Meier, R., & Dikow, T. (2004). Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, 18, 478–488. <https://doi.org/10.1111/j.1523-1739.2004.00233.x>
- Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 19, 992–1006. <https://doi.org/10.1111/ele.12624>
- Nicolson, N. (2019). *Automating the construction of higher order data representations from heterogeneous biodiversity datasets* [Dissertation, Brunel University London].
- Otegui, J., Ariño, A. H., Encinas, M. A., & Pando, F. (2013). Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS One*, 8, e55144. <https://doi.org/10.1371/journal.pone.0055144>
- Pebesma, E., & Bivand, R. S. (2005). S classes and methods for spatial data: The sp package. *R News*, 5, 9–13. <https://CRAN.R-project.org/doc/Rnews/>
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.R-project.org/>
- Raes, N., & Aguirre-Gutiérrez, J. (2018). Modeling framework to estimate and project species distributions space and time. In *Mountains, climate and biodiversity* (Vol. 309). John Wiley & Sons.
- Reydon, T. A., & Kunz, W. (2019). Species as natural entities, instrumental units and ranked taxa: New perspectives on the grouping and ranking problems. *Biological Journal of the Linnean Society*, 126, 623–636. <https://doi.org/10.1093/biolinnean/blz013>
- Schmidt-Lebuhn, A. N., Knerr, N. J., & Kessler, M. (2013). Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation*, 22, 905–919. <https://doi.org/10.1007/s10531-013-0457-9>
- Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2015). Package ROCR. Visualizing the performance of scoring classifiers. *Bioinformatics*, 21, 3940–3941. <https://doi.org/10.1093/bioinformatics/bti623>
- Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359, 689–698. <https://doi.org/10.1098/rstb.2003.1439>
- Sterner, B., & Franz, N. M. (2017). Taxonomy for humans or computers? Cognitive pragmatics for big data. *Biological Theory*, 12, 99–111. <https://doi.org/10.1007/s13752-017-0259-5>
- Stevenson, D. W. (1993). In *Flora of North America* Editorial Committee (Ed.), *Flora of North America*, volume 2: *Ephedraceae* (pp. 428–434). Oxford University Press. http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=10313
- Stropp, J., Ladle, R. J., M. Malhado, A. C., Hortal, J., Gaffuri, J., H. Temperley, W., Olav Skøien, J., & Mayaux, P. (2016). Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography*, 25, 1085–1096. <https://doi.org/10.1111/geb.12468>
- Tessarolo, G., Ladle, R., Rangel, T., & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, 7, 6863–6870. <https://doi.org/10.1002/ece3.3259>
- Thiers, B. (2022). (Continuously updated). *Index Herbarium: A global directory of public herbaria and associated staff*. New York Botanical Garden's Virtual Herbarium. [Internet]. <http://sweetgum.nybg.org/science/ih>
- Thuiller, W., & Lafourcade, B. (2019). *biomod2* package, pseudo.abs function. <https://rdrr.io/rforge/BIOMOD/man/pseudo.abs.html>
- Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). Package 'biomod2'. Species distribution modeling within an ensemble forecasting framework. <https://CRAN.R-project.org/package=biomod2>
- Töpel, M., Zizka, A., Calió, M. F., Scharn, R., Silvestro, D., & Antonelli, A. (2017). *SpeciesGeoCoder*: Fast categorization of species occurrences for analyses of biodiversity, biogeography, ecology, and evolution. *Systematic Biology*, 66, 145–151. <https://doi.org/10.1093/sysbio/syw064>
- Vetter, M. (1990). *Aufbau betrieblicher Informationssysteme mittels konzeptioneller Datenmodellierung*. 5. Auflage. Springer Verlag.
- Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). Package 'corrplot'. *Statistician*, 56(316), e24.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software*, 4, 1686. <https://doi.org/10.21105/joss.01686>
- Wiczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, 7, e29715. <https://doi.org/10.1371/journal.pone.0029715>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international*

conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29–40). <http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>

- Wortley, A. H., & Scotland, R. W. (2004). Synonymy, sampling and seed plant numbers. *Taxon*, 53, 478–480. <https://doi.org/10.2307/4135625>
- Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the global biodiversity information facility? *PLoS One*, 2, e1124. <https://doi.org/10.1371/journal.pone.0001124>
- Zermoglio, P. F., Guralnick, R. P., & Wieczorek, J. R. (2016). A standardized reference data set for vertebrate taxon name resolution. *PLoS One*, 11, e0146894. <https://doi.org/10.1371/journal.pone.0146894>
- Zhang, Y., Schaap, M. G., & Zha, Y. (2018). A high-resolution global map of soil hydraulic properties produced by a hierarchical parameterization of a physically based water retention model. *Water Resources Research*, 54, 9774–9790. <https://doi.org/10.1029/2018WR023539>
- Zizka, A. (2019). *Cleaning GBIF data for the use in biogeography (Tutorial)*. https://ropensci.github.io/CoordinateCleaner/articles/Cleaning_GBIF_data_with_CoordinateCleaner.html
- Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R., Colli-Silva, M., Fantinati, M. R., Fernandes, M. F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N. M. C., Santos, T. A. B., dos Santos-Costa, R. C., Serrano, F. C., Alves da Silva, A. P., de Souza Soares, A., Cavalcante de Souza, P. G., Calisto Tomaz, E., ... Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *Biodiversity and Conservation*, 8, e9916. <https://doi.org/10.7717/peerj.9916>
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Antonelli, A. (2019). *CoordinateCleaner*: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10, 744–751. <https://doi.org/10.1111/2041-210X.13152>

SUPPORTING INFORMATION

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

How to cite this article: Fühding-Potschkat, P., Kreft, H., & Ickert-Bond, S. M. (2022). Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models. *Ecology and Evolution*, 12, e9168. <https://doi.org/10.1002/ece3.9168>