

## Data Paper

# Predicted occurrence and abundance habitat suitability of invasive plants in the contiguous United States: updates for the INHABIT web tool

Catherine S. Jarnevich<sup>1</sup>, Peder Engelstad<sup>2</sup>, Demetra Williams<sup>1</sup>, Keana Shadwell<sup>1</sup>,  
Cameron Reimer<sup>1</sup>, Grace Henderson<sup>1</sup>, Janet S. Prevey<sup>1</sup>, Ian S. Pearse<sup>1</sup>

<sup>1</sup> U.S. Geological Survey, Fort Collins Science Center, Fort Collins, USA

<sup>2</sup> Graduate Degree Program in Ecology, Colorado State University in cooperation with the U.S. Geological Survey, Fort Collins Science Center, Fort Collins, USA

Corresponding author: Catherine S. Jarnevich (jarnevichc@usgs.gov)

## Abstract

Invasive plant species have substantial negative ecological and economic impacts. Geographic information on the potential and actual distributions of invasive plants is critical for their effective management. For many regions, numerous sources of predictive geographic information exist for invasive plants, often in the form of outputs from species distribution models (SDMs). The creation of a repository of consistently produced SDMs of regional- or national-scale information predicting the potential distribution of invasive plant species could provide information to managers in the prioritisation of invasive species management. Here, we present a novel set of not only habitat suitability models for occurrence for 259 manager requested invasive plant species in the contiguous United States (USA), but also habitat suitability models for abundance ( $\geq 5\%$  cover) and high abundance ( $\geq 25\%$  cover). These data provide an update to the Invasive Species Habitat Tool (INHABIT; gis.usgs.gov/inhabit). This tool contains information on the majority of invasive plant species in the contiguous USA with sufficient location data for model building. INHABIT provides a canonical set of predicted geographic distributions for invasive plants in the contiguous USA that can aid in the search for new populations of invasive plant species and help create watch lists for emerging invaders. As this tool contains information on nearly all of the most problematic invasive plants in the contiguous USA, it helps in prioritising management strategies by showing which plants are already present or abundant in a land management area and which may become present or abundant in the future.

**Key words:** Early detection and rapid response, land management, species distribution modelling, watch list



**Academic editor:** Ramiro Bustamante

**Received:** 21 August 2024

**Accepted:** 31 October 2024

**Published:** 25 November 2024

**Citation:** Jarnevich CS, Engelstad P, Williams D, Shadwell K, Reimer C, Henderson G, Prevey JS, Pearse IS (2024) Predicted occurrence and abundance habitat suitability of invasive plants in the contiguous United States: updates for the INHABIT web tool. NeoBiota 96: 261–278. <https://doi.org/10.3897/neobiota.96.134842>

**Copyright:** This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

## Introduction

Invasive species cause considerable damage to ecological and economic systems worldwide (Bellard et al. 2016; Diagne et al. 2021; Mayfield et al. 2021). In the United States of America (USA), this results in an annual cost exceeding \$19 billion per year (Fantle-Lepczyk et al. 2022). The effective management of invasive plants relies on information about where particular plant species are present and abundant (Wallace and Barger 2022) and where they are likely to become established and abundant in the future (Mainali et al. 2015). Proactive management of invasive species is often considered the most cost-effective strategy for control,

but this strategy is likely underemployed (Cuthbert et al. 2022), in part due to the lack of predictive tools and data to guide searches for invasive species and the prioritisation of management actions.

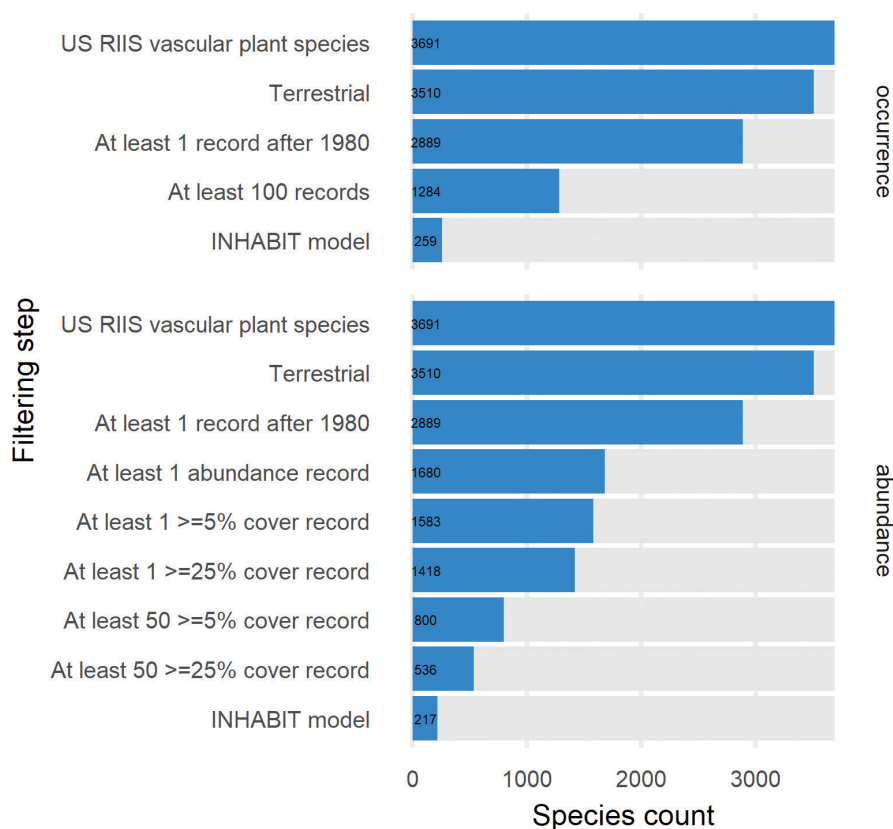
Predictive habitat mapping, based on species distribution models (SDMs), is the primary tool used to anticipate where invasive plant species will establish and become abundant (Crall et al. 2013; Elith 2017). SDMs can be useful in guiding management if the models used meet the standards needed for particular management objectives (Sofaer et al. 2019). Often, the lack of consistent methods, species or geographic scope can make it difficult to use or compare disparate models to prioritise the management and detection of invasive plant species. While there is often a perceived trade-off between the geographic scope and precision with SDMs, evidence suggests otherwise. Even at local scales, range-wide models often outperform ecoregion extent models of predicted species distributions, likely due to larger datasets that represent a species' niche more accurately (Jarnevich et al. 2022). The practical value of publicly available, high-quality SDMs across large spatial extents, such as the contiguous USA and at a fine spatial grain (~ 100 m<sup>2</sup>) for a comprehensive set of invasive plant species, is widely recognised amongst invasive species managers (Engelstad et al. 2022).

Most SDMs predict the habitats and geographic locations in which an invasive species might establish (i.e. become present). However, managers also want to know where species might become abundant as abundance is correlated with impact (Sofaer et al. 2018; Bradley et al. 2019; Pearse et al. 2019). To address this, SDMs can be fitted with sites where an invasive plant species is known to reach some threshold of abundance (e.g. over 25% cover; Jarnevich et al. (2021); Beaury et al. (2023); Evans et al. (2024)). Some management strategies require information about different thresholds of abundance to prioritise the application of control methods (Yokomizo et al. 2009). For example, areas of the Great Basin are twice as likely to burn when exceeding a threshold of 15% cover of invasive cheatgrass (*Bromus tectorum*) compared to habitats where cheatgrass is less abundant (Bradley et al. 2018). Additionally, while cheatgrass occurs in all 50 States, it is only in the western USA where it reaches a level of abundance that makes it a species of concern. Management strategies that focus on cheatgrass control for fire mitigation may choose to concentrate efforts on regions exceeding this threshold for cheatgrass cover. While many management actions rely on information about the abundance of an invasive plant species, information on where the species is likely to establish, even at low abundances, is critical for tracking dispersal. This is especially true for early detection and rapid response (EDRR) strategies that focus on the control of non-native species before they become abundant. Likewise, control of non-abundant invasive plant species may be a desirable proactive measure because the geographic locations of habitats in which an invasive plant species may become abundant are likely shifting due to changing climate (Evans et al. 2024).

Effective management of invasive plant species requires the prioritisation of species to control (Kumschick et al. 2012). Globally, more than 13,000 plant species have been introduced and established outside of their native ranges (van Kleunen et al. 2015). Within the USA, many of these introduced species do not become problematic and exist within the novel range with relatively few negative consequences (Bradley et al. 2024). However, a percentage of those introduced plant species become invasive, which we define as a non-native species that cause harm to the environment, economy or human, animal or plant health. Comprehensive

information about all invasive plant species in a region can aid in prioritising management decisions. For example, managers of particular land management areas are especially interested in those invasive plant species that have not yet been detected within their management area, but which have the potential to establish there (Jarnevich et al. 2023a). We developed the Invasive Species Habitat Tool (INHABIT; [gis.usgs.gov/inhabit](https://gis.usgs.gov/inhabit)) to provide a consistent, comprehensive set of habitat suitability models for a wide variety of invasive plant species to provide this information to managers (Jarnevich et al. 2023a).

In this data paper, we present the first abundance-based suitability models for a large suite of plant species through version 4 of INHABIT and describe the methodology used in its creation. This dataset is the first publicly available resource to provide a large number of predicted habitat suitability maps and management area summary tables for the habitats in which invasive plants may establish and may become abundant. The scope of the tool is invasive terrestrial plant species in the contiguous USA and includes 23% of all introduced vascular plant species with at least 100 georeferenced records and 50% of those with high abundance records (Fig. 1). As the inclusion of plant species in INHABIT was based on requests from diverse invasive species managers over a period of five years, it includes



**Figure 1.** The total number of non-native vascular plant species on the US Register for Introduced and Invasive Species (RIIS; Simpson et al. 2022) with different filtering steps applied for occurrence (top) and abundance (bottom) records, with the count of species remaining at each step reported. Filtering steps include removing species flagged as aquatic based on inclusion in species tracked in the U.S. Geological Survey's Non-Indigenous Aquatic Species (NAS; Terrestrial) and followed by whether any records from the target background dataset from aggregated data sources have been collected after 1980. Occurrence record filtering includes number of species with at least 100 occurrence records post 1980. The abundance filtering includes successive criteria for abundance records for different percent cover thresholds and record counts. The last step for both is the number of remaining species modelled for INHABIT, demonstrating the much higher capture of non-native plant species in the USA with recorded abundance populations then the number of occurrence records or the total on the US RIIS list.

nearly all of the non-native plant species considered most problematic by managers within the contiguous USA. The tool uses a nested tabular design to serve management-relevant information on plant species' distributions and displays details, such as model performance, continuous suitability surfaces and environmental predictors of suitable habitat for each species. We additionally make suitability maps easily downloadable for user-selected species and provide simple methods for importing those maps into the platforms commonly used by land managers to guide searches for new populations of invasive plants (Suppl. material 1: supplement 2, Field Maps instruction). INHABIT provides the most comprehensive predicted habitat suitability information for both occurrence and abundance of invasive plants in the contiguous USA and serves as a blueprint for modelling and data delivery in other regions.

## Methods

The model fitting and summarisation methodology described below represents elements first described by Young et al. (2020), iterated by Engelstad et al. (2022) and used in version 3 of INHABIT by Jarnevich et al. (2023b). For version 4 of INHABIT, we have further updated the methods used to produce version 3. Differences between versions 3 and 4 include an increased number of species, updated occurrence records through 2023, spatial cross-validation of models and models of abundance and high abundance suitability along with occurrence suitability. All code is written in R version 4.4. (R Core Team 2024) and scripts are available in Suppl. material 1: supplement 3. Model inputs and outputs are available from Jarnevich et al. (2024) at <https://doi.org/10.5066/P14HNEJF>.

## Species data

We asked people managing invasive plant species within the contiguous USA to contribute to a list of terrestrial non-native plant species to include in version 4 of INHABIT. The resulting list identified 286 non-native species. We obtained species occurrence and abundance data for these selected species from existing aggregated occurrence and agency databases (Suppl. material 1: table S1), including: the Early Detection and Distribution Mapping System (EDDMapS), the Bureau of Land Management's (BLM) Assessment, Inventory and Management (AIM) and Lotic databases, the BLM and National Park Service's National Invasive Species Information Management System (NISIMS), a published aggregated plant dataset (Bradley et al. 2024), the Standardised Plant Community with Introduced Status (SPCIS) database (Petri et al. 2023), the LANDFIRE reference database and the Goldwater weeds database for both types of data and the Global Biodiversity Information Facility (GBIF) and iMapInvasives database for occurrence only (see Suppl. material 1: table S1 for additional details). We used the Integrated Taxonomic Information System (ITIS; [www.itis.gov](http://www.itis.gov)) as a taxonomic authority to obtain all synonyms for all species using the R package "taxizedb" (Chamberlain et al. 2023).

We restricted species records and observations based on multiple geographic and data quality criteria. We retained records with observation dates  $\geq 1980$ , listed as "observation" or "specimen only" observation types (GBIF only) and a coordinate uncertainty  $\leq 30$  m. Additionally, we used the "CoordinateCleaner" package

(Zizka et al. 2019) to flag and remove potentially erroneous records, including those located near country capitals, country centroids, GBIF headquarters and known biodiversity institutions (e.g. research centres, universities, herbaria, museums, zoos and botanical gardens), or coordinates over oceans or other potential data errors (e.g. latitude and longitudes equal to zero, latitudes and longitudes with identical values).

We also obtained locations for all vascular plant species included on the U.S. Register of Introduced and Invasive Species (US-RIIS) list ver. 2.0 (Simpson et al. 2022) using the same criteria for inclusion. These data were used to control for sampling biases as detailed in the modelling description below. The US-RIIS documents introduced (non-native) species that are established (reproducing). We filtered this list to those at the species level (removed hybrids and sub-species) and limited it to vascular plant species (phylum Tracheophyta). These data were separated into two files, one with all georeferenced observations for the selected US-RIIS species, and another further filtered to observations that contained abundance information for species within this list.

## Modelling group

The greatest advance for version 4 of the INHABIT webtool is the inclusion of SDMs that predict species abundance in addition to SDMs that predict occurrence. Through informal discussions with managers related to previous work with occurrence and abundance habitat suitability models (occurrence, abundance as > 10% cover; Jarnevich et al. (2021)) (occurrence and four abundance groups; Beaury et al. (2023)), we decided to include three groups of habitat suitability models in version 4: occurrence (or presence), abundance and high abundance. Manager feedback indicated the two categories were too few (Jarnevich et al. 2021) and the five categories too many (Beaury et al. 2023) and most people providing feedback desired a low abundance category (around 5%) and a category around 20%. We settled on data classification into occurrence, abundance ( $\geq 5\%$  cover) and high abundance ( $\geq 25\%$  cover) categories, based on that manager input and data availability (cover class bins commonly found in input datasets where 25% was a more commonly used cut-off than 20%; Suppl. material 1: fig. S1).

Each observation record was classified into high abundance ( $\geq 25\%$  cover), abundance ( $\geq 5\%$  - 25% cover) or occurrence ( $< 5\%$  cover or no abundance information). Where numerical cover data were provided, we assigned abundance categories to those records. When there were numerical bins, we used the minimum cover in the bin as a conservative match to our categories, so that 5–30% would be assigned to the abundance bin. However, some aggregated occurrence databases included qualitative descriptions of abundance, which we manually classified: occurrence = “trace”, “rare”, “sparse”, “single plant”, “spot”, “light”, “low”; abundance = “medium”, “moderate”, “common”, “patch”, “patchy”, “scattered dense patches”; high abundance = “high”, “dense”, “abundant”, “heavy”, “major”, “dense monoculture”, “dominant cover”.

We used a nested set of observation records to fit models. Occurrence suitability models included observation records for a species from all three categories. Abundance suitability models were fitted with occurrence records from both the abundance and high abundance categories ( $\geq 5\%$  cover). High abundance suitability models only used observations categorised as high abundance ( $\geq 25\%$  cover).

## Data preparation

We spatially thinned species records by reducing observations to a minimum 900 m distance between points using the “geoThin” function in the R package “enmSdmX” (Smith et al. 2023) to limit spatial autocorrelation while preferentially retaining points with the highest abundance cover class. Additionally, we removed records falling within waterbodies by using our percentage clay soils predictor as a mask because the soil layers did not have values in areas of water (Suppl. material 1: table S2).

We required at least 100 spatially thinned observations within the contiguous USA to generate an occurrence model for a species. Any requested species with fewer observations were flagged for future modelling efforts using globally sourced observations and predictor data. For abundance and high abundance models, we required at least 50 spatially thinned observations for each model group and we only considered fitting abundance models for species for which we could fit occurrence models. Through previous INHABIT iterations, we have found it difficult to fit models with less than 50 locations and, unlike occurrence data beyond the USA, we are unaware of global repositories that include abundance information.

In statistical and machine learning communities (Hastie et al. 2009; Lever et al. 2016; Kuhn and Johnson 2019), best practices for model building advocate splitting data into three tiers including train, test and sub-setting train into k-fold cross-validation splits. The train data are used to fit the model, with the cross-validation subsets providing information for model refinement. The test partition is used to evaluate model performance as this partition is fully withheld from the model fitting and refinement process. We implemented a hashtag-based method to split each species’ data into non-random, spatially sampled cross-validation (CV) data for model training and a separate withheld test dataset to evaluate model performance. This methodology uses a hashtag shape (#) overlaid on a 99% binary kernel density estimate (KDE) of the observations. Observation points falling within the buffered hashtag were assigned to the test split (and withheld from model fitting to be used for model evaluation) with a desired ratio of 70% train/30% test. The width of the hashtag test strips was manually adjusted from the default of 30% of the binary KDE extent on a case-by-case basis as needed to stay within the range of 20–50% of observations in the test split (ideally around 30%). The training data were further split spatially, based on the hashtag shape into nine CV splits. We used a single hashtag shape to define the spatial splits (both test and CV) for all three model groups and, thus, the test data for abundance were a subset of occurrence and for high abundance were a subset of abundance. We did not include a test split in the few cases where that would result in < 100 observations to train the occurrence model ( $n = 5$  species, Suppl. material 1: table S3), though these still had a CV split. In cases of high geographic clustering resulting in algorithms failing to fit models with the spatial CV splits, we instead applied a randomised 9-fold CV split (i.e. no spatial CV-split).

As we did not have absence data, we required background locations to capture the environments available to each species to fit the models. We used two methods to generate background points for occurrence model training data to fit two sets of models for each species to account for sampling biases, a continuous KDE method and a target background approach. The continuous KDE method has been suggested for invasive species in particular because there may be a higher density of observations in a region to which the species has been introduced longer compared to the density of observations in a region where a species has only arrived recently

(Elith et al. 2010), whereas the target background approach is meant to mimic bias in where people are collecting observations (Phillips et al. 2009). For the KDE method, we created a continuous KDE raster around the observations that were used to weight the generation of an initial 15,000 random background points. We spatially thinned the background points in the same manner as the occurrence records, randomly selecting 10,000 of the remaining points. For the target background method, we randomly selected up to 10,000 locations of non-native vascular plant observations (from the US-RIIS list) within the matching lifeform of four possible lifeforms assigned by the USDA Plants Database (forb/herb/vine, graminoid, shrub/vine or tree) and restricted to the same 99% binary KDE used for the hashtag splits. As there is a relationship between residence time and abundance, we chose to only use the target background approach for the abundance models, assuming that these models will suffer less from issues related to sampling density due to residence time and more from sampling bias related to where people are making observations. We selected background points from the abundance observations of the non-native plants from within the same broad lifeform category. For all three model groups, training data background points were excluded from the buffered hashtag-defined test extent and were assigned to CV splits using the same spatial designation as for the occurrence/abundance records employed for the individual species as described above. For test splits, we generated one set of background points by rasterising the buffered hashtag shape and subsampling 10,000 randomly distributed points separated by at least 900 metres from within this shape. Thus, all model groups had the same background test points for each species.

## Environmental data

We used 52 of the 54 environmental predictors included in INHABIT version 3 (Jarnevich et al. 2023b), representing a range of environmental factors including topography, temperature, atmospheric water, landscape water, soil properties, disturbance (including fire and anthropogenic disturbance), biotic interactions (e.g. tree cover and bare ground) and radiation for the contiguous USA. These predictors were used as inputs for version 4 models after they were modified using the PARC (Project, Aggregate, Resample, Clip) module in the Software for Assisted Habitat Modelling (SAHM, v. 2.2.2; Morissette et al. 2013) to ensure all predictors were in the same coordinate reference system (ESRI:102008), spatial resolution ( $\sim 100 \text{ m}^2$ ), extent and alignment. Observation data for each species were combined with predictor data to create a merged dataset representing the environmental conditions at each presence and background location used as an input in SAHM. For more information on individual predictors, their units of measure and spatial and temporal resolution, see Suppl. material 1: table S2.

For each species, we selected predictors based on individual species characteristics including biology and lifeform (e.g. winter annual graminoid) and invaded geographic distribution within the contiguous USA. Predictor sets were consistent between occurrence models (i.e. between the KDE background approach and target background approach) and were kept identical between abundance and high abundance models, except in a few cases where the number of predictors for high abundance were highly restricted based on number of observations. To avoid autocorrelation amongst potential predictors, we assessed the degree of correlation using the Pearson, Spearman and Kendall's pairwise correlation tests and removing one of

any pair with a correlation coefficient  $> 0.70$  (maximum of Pearson, Spearman or Kendall; Dormann et al. (2013)). We also removed any predictor that did not make ecological sense for a particular species (e.g. ratio of March precipitation for a tree species occurring in the eastern USA). In addition, we preferentially retained predictors known to be important for the distribution of invasive plants in the contiguous USA, specifically minimum winter temperature and index of human modification (Williams et al. 2024). We maintained a ratio of at least 10:1 observations to predictors (Hosmer and Lemeshow 2000), rounding up to the nearest ten. This resulted in a range of 10 to 29 predictors (mean 23.8 predictors) for each species for model fitting for occurrence models and a range of 5–31 predictors (mean 17.9 predictors) for each species for model fitting for abundance and high abundance models.

### Model fitting

Following the methodology in Young et al. (2020) and Engelstad et al. (2022), we fit five different algorithms in SAHM including boosted regression trees (Elith et al. 2008), generalised linear models (McCullagh and Nelder 1989), multivariate adaptive regression splines (Elith and Leathwick 2007), Maxent (Phillips et al. 2017) and random forests (Breiman 2001). We updated the SAHM downsampling code for random forests to balance the number of bootstrapped samples from each class (presence and background) and improved model performance over default settings (Valavi et al. 2022). We fit models with observation training data delineated by the hashtag (see Suppl. material 1: table S3 for individual sample sizes and number of cross-validation splits). We examined each algorithm's output, using *a priori* criteria to identify model overfitting, including instances when the difference between the Area Under the Curve (AUC) calculated for training data and the average cross-validation AUC was  $> 0.05$  or visual assessment of response curves appearing overly complex. In these cases, we explored alternative algorithm-specific tuning parameters to decrease the AUC difference and response curve complexity (see Suppl. material 1: table S3 for individual model changes). Finally, we assessed model performance using the continuous Boyce index (CBI; Hirzel et al. (2006)) calculated for each algorithm for each model per species and the AUC, revisiting any model where train or test CBI  $< 0.5$  or AUC  $< 0.7$  to reconsider parameterisation and dropping models with poor performance and fit.

### Spatial outputs: maps and ensembles

For each model group, we created continuous spatial predictions of relative habitat suitability across the contiguous USA at  $\sim 100 \text{ m}^2$  spatial resolution, expedited by U.S. Geological Survey high performance computing resources (Falgout et al. 2024). We also calculated a multivariate environmental similarities surface (MESS) to highlight areas of model extrapolation that could be applied as a mask to any output maps (Elith et al. 2010). The MESS map compares the value for each predictor at a location to the range of values for the predictor within the training set and we highlight locations with negative values that indicate at least one predictor included in a model set had a value outside those of the training data, termed novel environmental conditions. As we had background points rather than absence data, models were not necessarily calibrated to each other. Therefore, we rescaled the mapped values for each model between 0 and 100 to make the maps

more comparable. We then produced ensemble maps for each model group by taking a mean of the rescaled relative habitat suitability values, weighted by individual model CBI values (using the test CBI, if available). Individual models with CBI values of less than 0.50 were assessed and dropped if deemed appropriate, based on examining models for the specific species considering how low CBI was, how it compared to other algorithm predictions and ecological plausibility of response curves and variable importance (see notes column in Suppl. material 1: table S3 for specific reasons for retaining or dropping individual algorithms).

Through informal discussions with managers during presentations, managers expressed interest in having categorical maps of suitability along with continuous relative predictions of suitability for the three model groups. They reported an interest in three options as existed on INHABIT version 3, ranging from more inclusive to more restrictive. Thus, we produced three distinct binary versions corresponding to each of the three continuous maps (occurrence, abundance, high abundance) using percentile thresholds of 1%, 5% and 10% to convey a gradient of inclusive (comprehensive) to restrictive (targeted) model output (Engelstad et al. 2022). A percentile threshold is calculated by extracting predicted values for all training observations and selecting the predicted value for the observation that would misclassify the set percentage of positive observations. Finally, we combined the binary maps to display information across all three model groups (occurrence, abundance, high abundance) for each of the three thresholds, while highlighting any areas of environmental extrapolation.

### **Tabular outputs: zonal summaries and distance measurements**

Managers can utilise summaries of habitat suitability for management areas in various ways, such as to create watch lists (e.g. Jarnevich et al. 2023a) or find out how many management areas have suitable habitat for a particular species. INHABIT version 3 included tables summarising mapped outputs to quantify area of suitability and distance to closest records for some federal lands (Bureau of Land Management, Fish and Wildlife Service, National Park Service, U.S. Forest Service) and U.S. Counties. For INHABIT version 4, we obtained 10,216 management area polygons to meet stakeholder, manager and agency needs including additional federally owned lands (U.S. Army Corps of Engineers, Department of Defense service branches), State Parks and Forests, sub-basins (8-digit hydrologic unit code) and other more specific land management areas, such as cooperative invasive species management areas. Descriptions of the management areas and the additional processing steps necessary to prepare them for analyses are detailed in Suppl. material 1: table S4.

We summarised the categorical maps for management areas (Suppl. material 1: table S4) using the “exactextractr” R package (Baston 2023). For each species and management area, we calculated the total land area of suitable habitat for each of the three model groups (occurrence, abundance, high abundance) for each of the three thresholds (1%, 5%, 10%). Values were nested such that summaries of occurrence models included locations defined as suitable by any of the three model groups. We also calculated the total area that was identified by the MESS maps as having novel environmental conditions (areas of extrapolation as identified by any single predictor having a value outside the range of the model training data for that predictor). To provide further information on uncertainty, we calculated the area of a management area that was available for modelling, removing the parts of

the management area that we were unable to make predictions for which typically represented water, but potentially other locations where all predictors did not have data (though this is exceedingly rare).

We also generated an updated dataset containing the modelled species' observation locations to capture the most recent observations following the same steps outlined in the species data section above. Using these data, we counted the number of occurrence locations within each management area boundary and measured the distance from boundaries to the nearest location when no observations fell within the management area.

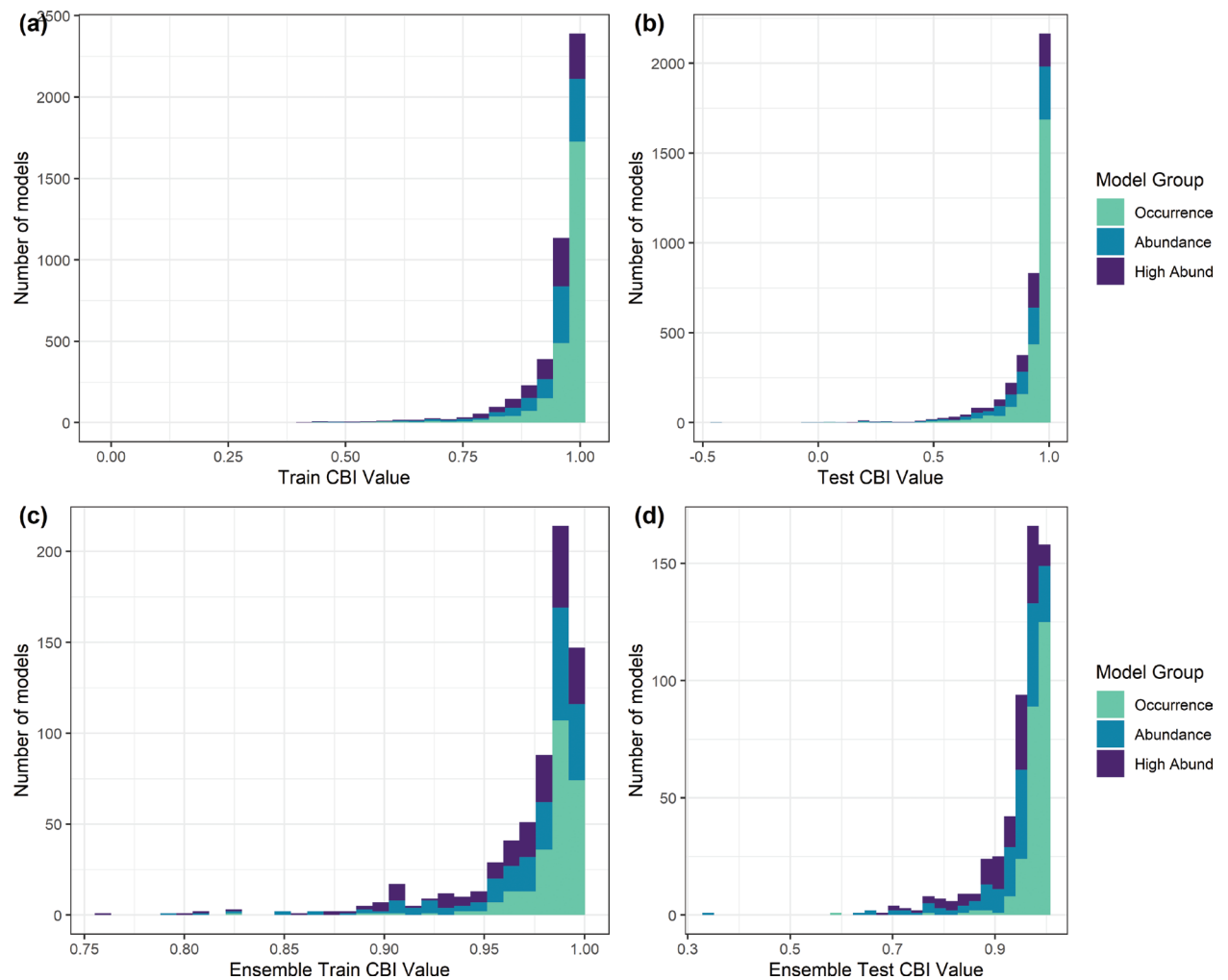
We merged the habitat suitability summary information with the count and distance information to provide information for watch list development. Early detection at a local level can be informed by watch lists of doorstep invaders, which we define as species with habitat suitability in the focal area, with no known records within the area and with records within either a 50- or 100-mile (75–150 km) buffer of the area (Jarnevich et al. 2023a). As early detection may be most effective in preventing establishment following secondary spread into a new region, we have added a new set of larger areas. We included summaries for ecoregions, watersheds and state boundaries that, when combined with other management areas, can identify species with suitable habitat within a management area which are not yet found within the larger region.

## Results

Managers requested models of 286 species. Of these 286 species, 254 species had at least 143 observations for a spatial test/train data split and five had at least 100 filtered observations for a model to be fitted with no test split. Twenty-seven species had < 100 filtered occurrences and, therefore, did not have models fit. Additionally, 217 species had at least 50 abundance observations ( $\geq 5\%$  cover) and 189 had at least 50 high abundance observations ( $\geq 25\%$  cover).

Overall, models performed well with most CBI values, based on the withheld test data,  $> 0.75$  for both individual algorithms and ensemble models for all three model groups (Fig. 2). There were 16 species that had at least one algorithm excluded from the ensemble due to poor occurrence model performance (typically  $\text{CBI} < 0.5$ ). This number was 25 for abundance models and 18 for high abundance models. This resulted in 4933 maps produced to develop the 665 continuous ensemble prediction maps and 259 categorical maps for each of the three thresholds. Based on an evaluation rubric for species distribution models, the models were categorised as acceptable or ideal (Suppl. material 1 table S5; Sofaer et al. 2019).

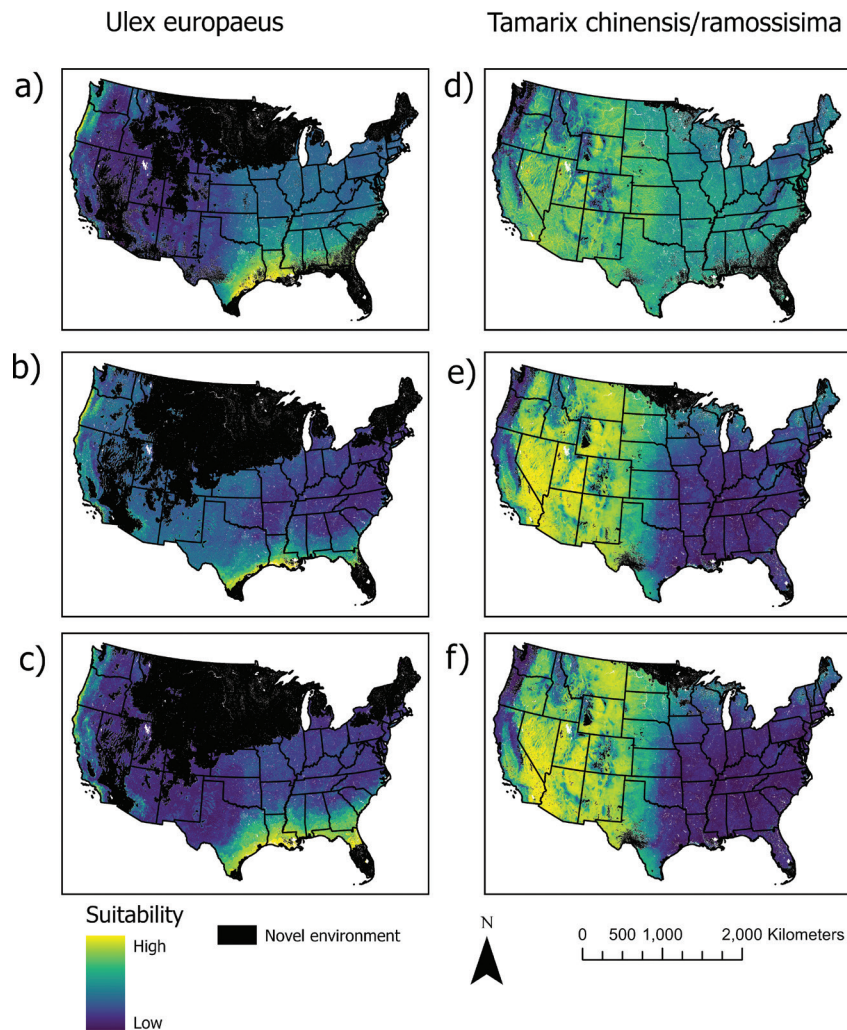
The continuous maps display relative habitat suitability at a  $\sim 100 \text{ m}^2$  resolution for the contiguous USA. There is a relative suitability map for each model group, including occurrence, abundance and high abundance (see examples in Fig. 3). While the example in Fig. 3 includes masking of areas of environmental extrapolation, both the available data and the webtool also include map versions without masking. Habitat suitability patterns differed between species. For example, *Ulex europaeus* (common gorse) had suitability concentrated in the southeast and northwest coast of the USA (Fig. 3a–c), while *Tamarix chinensis/ramosissima* (combination of two species of tamarisk that are difficult to distinguish and hybridise) had suitability concentrated in the west (Fig. 3d–f). *Tamarix chinensis/ramosissima* also had a large, pronounced difference in suitability for occurrence versus abundance or high abundance in the eastern USA.



**Figure 2.** Histograms of the continuous Boyce Index (CBI) for all fit models across algorithms and species calculated for the **a** training data **b** withheld test data and for the continuous ensemble models for the **c** training data and **d** withheld test data.

The integrated maps illustrated the differences in the thresholds used to generate them and showed differences in patterns between species (Fig. 4). By definition, the 1<sup>st</sup> percentile maps (Fig. 4a–d) contained a larger area of suitable habitat for each group than the 10<sup>th</sup> percentile maps (Fig. 4c, f). For our example species, *Ulex europaeus* showed a marked decrease in occurrence suitability and even the most targeted threshold contained visually discernible amounts in all categories of suitability (Fig. 3a–c). However, *T. chinensis/ramosissima* had much less occurrence suitability across all three maps, with areas of suitability for high abundance covering the largest area, indicating that, in most areas where the species could establish, it could also become abundant (Fig. 4d–f).

From the tabular summaries, across all species and management areas with suitability, the mean percentage suitable area ranged from 20% to 46% for occurrence, 11% to 26% for abundance and 7% to 19% for high abundance. See management summary tables in Jarnevich et al. (2024). The lower numbers correspond to data from the most restrictive threshold (“first”; derived from maps such as Fig. 3a, d) and the higher numbers correspond to the most inclusive threshold (“tenth”; derived from maps such as Fig. 3c, f). Of the model summaries, for 10216 management areas, 41 lacked predictions for at least one species due to a lack of information for at least one predictor.

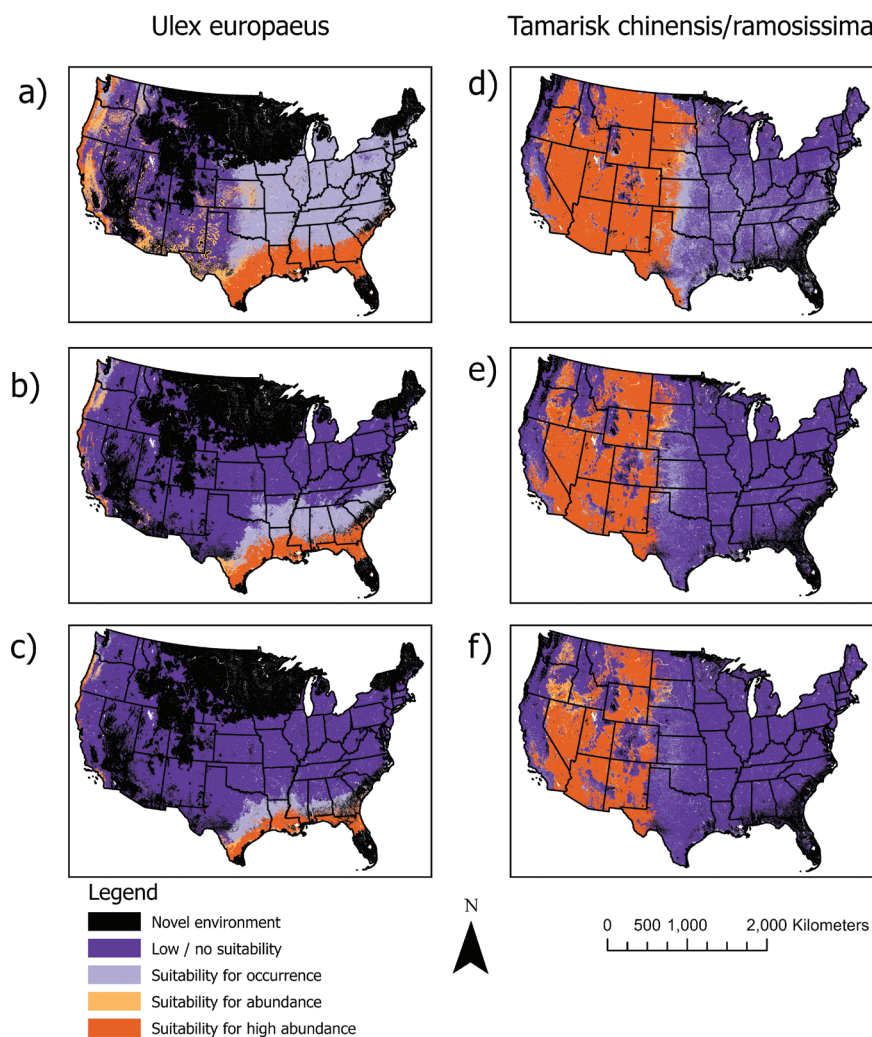


**Figure 3.** Continuous ensemble map for each of three model groups representing low to high habitat suitability for *Ulex europaeus* (a–c) and *Tamarix chinensis/ramossissima* (d–f) including a, d occurrence b, e abundance ( $\geq 5\%$  cover) and c, f high abundance ( $\geq 25\%$  cover). Black indicates areas with novel environmental conditions (values for at least one model predictor outside the range of values captured by the training data for model fitting).

## Discussion

We created updated models of habitat suitability for occurrence for 220 species, new occurrence models for 39 species and models of abundance habitat suitability for 217 species (Suppl. material 1: table S3, Fig. 1). These models are currently fitted using binary classification of abundance and demonstrate the potential for invasive species to reach high abundance and, hence, become problematic in new areas. As higher quality continuous abundance data become available across larger spatial extents, future models may be created showing continuous predictions for invasive plant abundance (e.g. Sofaer et al. 2022). Information regarding critical cover thresholds relating to impacts is often not available for individual invasive plant species, thus, more research in this area would improve abundance threshold selection and better inform managers about the ecological threats invasive plants pose at different levels of abundance.

Model applications for management and decision-making relevant to invasive plants are diverse. Regional early detection and rapid response applications allow for newly-introduced or actively expanding invasive plant species to be monitored prior to establishment in a management area. Similarly, an “invaders at the doorstep” approach uses



**Figure 4.** Composite categorical map from discretising the continuous ensemble maps in Fig. 2 using three different threshold values to calculate the combined binary maps including **a, d** 1<sup>st</sup> percentile **b, e** 5<sup>th</sup> percentile and **c, f** 10<sup>th</sup> percentile.

models to develop watch lists that could provide information for early detection efforts for species found nearby, but not yet within a management area. Abundance models can be similarly used for management applications, but can be applied to further refine surveys or control efforts to prioritise those species that may be more impactful in an area. When used in conjunction with spatial data on vegetation and wildlife resources, managers may better identify intersections between areas that support greater invasive plant abundance and areas with particularly vulnerable native communities. Managers may choose to prioritise actions based on model outputs and landscape features, such as when an area that is predicted to support higher abundance of an invasive plant is positioned next to a road or waterway that may further spread propagules.

## Additional information

### Conflict of interest

The authors have declared that no competing interests exist.

### Ethical statement

No ethical statement was reported.

## Funding

Funding for this project came from Bipartisan Infrastructure Law: Ecosystem Restoration Activity 6: Invasive Species and contributes this work to the National Early Detection and Rapid Response Framework. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

## Author contributions

Conceptualization: All; Data curation: CSJ, PE, DW, KS, CR; Formal analysis: all; Funding acquisition: CSJ, JSP, ISP; Investigation: all, Methodology: all, Project administration: CSJ, PE; Software: PE, DW, KS, CR, GH; Resources: CSJ, Supervision: CSJ; Validation: CSJ, PE, DW, KS; Visualization: PE, DW, KS, CR; Writing – original draft: all; Writing – review & editing: all.

## Author ORCIDs

Catherine S. Jarnevich  <https://orcid.org/0000-0002-9699-2336>

Peder Engelstad  <https://orcid.org/0000-0002-3681-9216>

Demetra Williams  <https://orcid.org/0000-0002-5171-8640>

Keana Shadwell  <https://orcid.org/0000-0001-6835-425X>

Cameron Reimer  <https://orcid.org/0000-0002-2058-0538>

Grace Henderson  <https://orcid.org/0000-0001-9542-6888>

Janet S. Prevey  <https://orcid.org/0000-0003-2879-6453>

Ian S. Pearse  <https://orcid.org/0000-0001-7098-0495>

## Data availability

All of the data that support the findings of this study and the study outputs are available in the main text, Supplementary Information or as a U.S. Geological Survey data release (Jarnevich et al. 2024; <https://doi.org/10.5066/P92476V6>).

## References

- Baston D (2023) exactextractr: Fast Extraction from Raster Datasets using Polygons. R package version 0.10.0. <https://github.com/isciences/exactextractr>
- Beaury EM, Jarnevich CS, Pearse I, Evans AE, Teich N, Engelstad P, LaRoe J, Bradley BA (2023) Modeling habitat suitability across different levels of invasive plant abundance. *Biological Invasions* 25(11): 3471–3483. <https://doi.org/10.1007/s10530-023-03118-z>
- Bellard C, Cassey P, Blackburn TM (2016) Alien species as a driver of recent extinctions. *Biology Letters* 12(2): 20150623. <https://doi.org/10.1098/rsbl.2015.0623>
- Bradley BA, Curtis CA, Fusco EJ, Abatzoglou JT, Balch JK, Dadashi S, Tuanmu M-N (2018) Cheatgrass (*Bromus tectorum*) distribution in the intermountain Western United States and its relationship to fire frequency, seasonality, and ignitions. *Biological Invasions* 20(6): 1493–1506. <https://doi.org/10.1007/s10530-017-1641-8>
- Bradley BA, Laginhas BB, Whitlock R, Allen JM, Bates AE, Bernatchez G, Diez JM, Early R, Lenoir J, Vilà M, Sorte CJB (2019) Disentangling the abundance-impact relationship for invasive species. *Proceedings of the National Academy of Sciences of the United States of America* 116(20): 9919–9924. <https://doi.org/10.1073/pnas.1818081116>
- Bradley BA, Evans AE, Jarnevich CS, Beaury EM, Engelstad P, Teich NB, LaRoe JM (2024) US non-native plant occurrence and abundance data and distribution maps for Eastern US species with current and future climate: U.S. Geological Survey data release.
- Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32. <https://doi.org/10.1023/A:1010933404324>

- Chamberlain S, Arendsee Z, Stirling T (2023) taxizedb: Tools for Working with ‘Taxonomic’ Databases. R package version 0.3.1.
- Crall AW, Jarnevič CS, Panke B, Young N, Renz M, Morissette J (2013) Using habitat suitability models to target invasive plant species surveys. *Ecological Applications* 23(1): 60–72. <https://doi.org/10.1890/12-0465.1>
- Cuthbert RN, Diagne C, Hudgins EJ, Turbelin A, Ahmed DA, Albert C, Bodey TW, Briski E, Essl F, Haubrock PJ, Gozlan RE, Kirichenko N, Kourantidou M, Kramer AM, Courchamp F (2022) Biological invasion costs reveal insufficient proactive management worldwide. *The Science of the Total Environment* 819: 153404. <https://doi.org/10.1016/j.scitotenv.2022.153404>
- Diagne C, Leroy B, Vaissière A-C, Gozlan RE, Roiz D, Jarić I, Salles J-M, Bradshaw CJA, Courchamp F (2021) High and rising economic costs of biological invasions worldwide. *Nature* 592(7855): 571–576. <https://doi.org/10.1038/s41586-021-03405-6>
- Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 027–046. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Elith J (2017) Predicting distributions of invasive species. In: Robinson AP, Walshe T, Burgman MA, Nunn M (Eds) *Invasive species: risk assessment and management* Cambridge University Press, 93–129. <https://doi.org/10.1017/9781139019606.006>
- Elith J, Leathwick J (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity & Distributions* 13(3): 265–275. <https://doi.org/10.1111/j.1472-4642.2007.00340.x>
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4): 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1(4): 330–342. <https://doi.org/10.1111/j.2041-210X.2010.00036.x>
- Engelstad P, Jarnevič CS, Hogan T, Sofaer HR, Pearse IS, Sieracki J, Frakes N, Sullivan J, Young NE, Prevey J, Belamaric P, LaRoe J (2022) INHABIT: A Web-Based Decision Support Tool for Invasive Species Habitat Visualization and Assessment Across the Contiguous United States. *PLoS ONE* 17(2): e0263056. <https://doi.org/10.1371/journal.pone.0263056>
- Evans AE, Jarnevič CS, Beaury EM, Engelstad PS, Teich NB, LaRoe JM, Bradley BA (2024) Shifting hotspots: Climate change projected to drive contractions and expansions of invasive plant abundance habitats. *Diversity & Distributions* 30(1): 41–54. <https://doi.org/10.1111/ddi.13787>
- Falgout JT, Gordon J, Lee L, Williams B (2024) USGS Advanced Research Computing, USGS Hovenweep Supercomputer: U.S. Geological Survey.
- Fantle-Lepczyk JE, Haubrock PJ, Kramer AM, Cuthbert RN, Turbelin AJ, Crystal-Ornelas R, Diagne C, Courchamp F (2022) Economic costs of biological invasions in the United States. *The Science of the Total Environment* 806: 151318. <https://doi.org/10.1016/j.scitotenv.2021.151318>
- Hastie T, Tibshirani R, Friedman J (2009) *Model Assessment and Selection. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer New York, New York, NY, 219–259. [https://doi.org/10.1007/978-0-387-84858-7\\_7](https://doi.org/10.1007/978-0-387-84858-7_7)
- Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* 199(2): 142–152. <https://doi.org/10.1016/j.ecolmodel.2006.05.017>
- Hosmer DW, Lemeshow S (2000) *Applied logistic regression.* Wiley, New York, 373 pp. <https://doi.org/10.1002/0471722146>
- Jarnevič CS, Sofaer HR, Engelstad P (2021) Modelling presence versus abundance for invasive species risk assessment. *Diversity & Distributions* 27(12): 2454–2464. <https://doi.org/10.1111/ddi.13414>

- Jarnevič CS, Sofaer HR, Belamaric P, Engelstad P (2022) Regional models do not outperform continental models for invasive species. *NeoBiota* 77: 1–22. <https://doi.org/10.3897/neobiota.77.86364>
- Jarnevič C, Engelstad P, LaRoe J, Hays B, Hogan T, Jirak J, Pearse I, Prevéy J, Sieracki J, Simpson A, Wenick J, Young N, Sofaer HR (2023a) Invaders at the doorstep: Using species distribution modeling to enhance invasive plant watch lists. *Ecological Informatics* 75: 101997. <https://doi.org/10.1016/j.ecoinf.2023.101997>
- Jarnevič CS, LaRoe J, Engelstad P, Hays B, Henderson G, Williams D, Shadwell K, Pearse IS, Prevéy JS, Sofaer HR (2023b) INHABIT species potential distribution across the contiguous United States (ver. 3.0, January 2023): U.S. Geological Survey data release. <https://doi.org/10.5066/P9V54H5K>
- Jarnevič CS, Engelstad P, Williams D, Shadwell K, Reimer C, Henderson G, Prevéy JS, Pearse IS (2024) INHABIT species potential distribution across the contiguous United States (ver. 4.0, June 2024): U.S. Geological Survey data release.
- Kuhn M, Johnson K (2019) Feature engineering and selection: A practical approach for predictive models. Chapman and Hall/CRC, New York, 310 pp. <https://doi.org/10.1201/9781315108230>
- Kumschick S, Bacher S, Dawson W, Heikkilä J, Sendek A, Pluess T, Robinson T, Kuhn I (2012) A conceptual framework for prioritization of invasive alien species for management according to their impact. *NeoBiota* 15: 69–100. <https://doi.org/10.3897/neobiota.15.3323>
- Lever J, Krzywinski M, Altman N (2016) Model selection and overfitting. *Nature Methods* 13(9): 703–704. <https://doi.org/10.1038/nmeth.3968>
- Mainali KP, Warren DL, Dhileepan K, McConnachie A, Strathie L, Hassan G, Karki D, Shrestha BB, Parmesan C (2015) Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. *Global Change Biology* 21(12): 4464–4480. <https://doi.org/10.1111/gcb.13038>
- Mayfield AE, Seybold SJ, Haag WR, Johnson MT, Kerns BK, Kilgo JC, Larkin DJ, Lucardi RD, Moltzan BD, Pearson DE, Rothlisberger JD, Schardt JD, Schwartz MK, Young MK (2021) Impacts of Invasive Species in Terrestrial and Aquatic Systems in the United States. *Invasive Species in Forests and Rangelands of the United States*. Springer International Publishing, 5–39. [https://doi.org/10.1007/978-3-030-45367-1\\_2](https://doi.org/10.1007/978-3-030-45367-1_2)
- McCullagh P, Nelder JA (1989) Generalized linear models, 2<sup>nd</sup> edn. Chapman and Hall, London; New York, 532 pp. <https://doi.org/10.1007/978-1-4899-3242-6>
- Morisette JT, Jarnevič CS, Holcombe TR, Talbert CB, Ignizio D, Talbert MK, Silva C, Koop D, Swanson A, Young NE (2013) VisTrails SAHM: Visualization and workflow management for species habitat modeling. *Ecography* 36(2): 129–135. <https://doi.org/10.1111/j.1600-0587.2012.07815.x>
- Pearse IS, Sofaer HR, Zaya DN, Spyreas G (2019) Non-native plants have greater impacts because of differing per-capita effects and nonlinear abundance-impact curves. *Ecology Letters* 22(8): 1214–1220. <https://doi.org/10.1111/ele.13284>
- Petri L, Beaury EM, Corbin J, Peach K, Sofaer H, Pearse IS, Early R, Barnett DT, Ibáñez I, Peet RK, Schafale M, Wentworth TR, Vanderhorst JP, Zaya DN, Spyreas G, Bradley BA (2023) SPCIS: Standardized Plant Community with Introduced Status database. *Ecology* 104(3): e3947. <https://doi.org/10.1002/ecy.3947>
- Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications* 19(1): 181–197. <https://doi.org/10.1890/07-2153.1>
- Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME (2017) Opening the black box: An open-source release of Maxent. *Ecography* 40(7): 887–893. <https://doi.org/10.1111/ecog.03049>

- R Core Team (2024) R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria.
- Simpson A, Fuller P, Faccenda K, Evenhuis N, Matsunaga J, Bowser M (2022) United States Register of Introduced and Invasive Species (US-RIIS) (ver. 2.0, November 2022): U.S. Geological Survey data release.
- Smith AB, Murphy SJ, Henderson D, Erickson KD (2023) Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography* 32(3): 342–355. <https://doi.org/10.1111/geb.13628>
- Sofaer HR, Jarnevič CS, Pearse IS (2018) The relationship between invader abundance and impact. *Ecosphere* 9(9): e02415. <https://doi.org/10.1002/ecs2.2415>
- Sofaer HR, Jarnevič CS, Pearse IS, Smyth RL, Auer S, Cook GL, Edwards Jr TC, Guala GF, Howard TG, Morissette JT, Hamilton H (2019) Development and delivery of species distribution models to inform decision-making. *Bioscience* 69(7): 544–557. <https://doi.org/10.1093/biosci/biz045>
- Sofaer HR, Jarnevič CS, Buchholtz EK, Cade BS, Abatzoglou JT, Aldridge CL, Comer PJ, Manier D, Parker LE, Heinrichs JA (2022) Potential cheatgrass abundance within lightly invaded areas of the Great Basin. *Landscape Ecology* 37(10): 2607–2618. <https://doi.org/10.1007/s10980-022-01487-9>
- Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J (2022) Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs* 92(1): e01486. <https://doi.org/10.1002/ecm.1486>
- van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, Kreft H, Weigelt P, Kartesz J, Nishino M, Antonova LA, Barcelona JF, Cabezas FJ, Cárdenas D, Cárdenas-Toro J, Castaño N, Chacón E, Chatelain C, Ebel AL, Figueiredo E, Fuentes N, Groom QJ, Henderson L, Inderjit, Kupriyanov A, Masciadri S, Meerman J, Morozova O, Moser D, Nickrent DL, Patzelt A, Pelser PB, Baptiste MP, Poopath M, Schulze M, Seebens H, Shu W, Thomas J, Velayos M, Wieringa JJ, Pyšek P (2015) Global exchange and accumulation of non - native plants. *Nature* 525(7567): 100–103. <https://doi.org/10.1038/nature14910>
- Wallace RD, Barger CT (2022) Identifying Invasive Species in Real Time: Early Detection and Distribution Mapping System (EDDMapS) and Other Mapping Tools. In: Ziska L (Ed.) *Invasive Species and Global Climate Change*. CABI, 225–238.
- Williams DA, Shadwell KS, Pearse IS, Prevéy JS, Engelstad P, Henderson GC, Jarnevič CS (2024) Predictor Importance in Habitat Suitability Models for Invasive Terrestrial Plants. *Diversity & Distributions* 13906(9): e13906. <https://doi.org/10.1111/ddi.13906>
- Yokomizo H, Possingham HP, Thomas MB, Buckley YM (2009) Managing the impact of invasive species: The value of knowing the density-impact curve. *Ecological Applications* 19(2): 376–386. <https://doi.org/10.1890/08-0442.1>
- Young NE, Jarnevič CS, Sofaer HR, Pearse I, Sullivan J, Engelstad P, Stohlgren TJ (2020) A modeling workflow that balances automation and human intervention to inform invasive plant management decisions at multiple spatial scales. *PLoS ONE* 15(3): e0229253. <https://doi.org/10.1371/journal.pone.0229253>
- 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(5): 744–751. <https://doi.org/10.1111/2041-210X.13152>

## Supplementary material 1

### Additional information

Authors: Catherine S. Jarnevich, Peder Engelstad, Demetra Williams, Keana Shadwell, Cameron Reimer, Grace Henderson, Janet S. Prevey, Ian S. Pearse

Data type: docx

Explanation note: **supplement 1.** Supplementary figures and tables. **supplement 2.** Field Maps instructions. **supplement 3.** Data processing R scripts.

Copyright notice: This dataset is made available under the Open Database License (<http://opendata-commons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/neobiota.96.134842.suppl1>