# Predicting hotspots for invasive species introduction in Europe

Kevin Schneider, David Makowski, Wopke van Der Werf

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

# Predicting hotspots for invasive species introduction in Europe

Kevin Schneider[1,*], David Makowski[2] and Wopke van der Werf[3]

[1] Business Economics Group, Wageningen University, PO Box 8130, 6700 EW Wageningen, The Netherlands
[2] Université Paris-Saclay, AgroParisTech, INRAE, Unit Applied Mathematics and Computer Science (MIA 518), Paris, France
[3] Centre for Crop Systems Analysis, Wageningen University, PO Box 430, 6700 AK Wageningen, The Netherlands
[*] Author to whom any correspondence should be addressed.

**E-mail:** k.schneider2304@gmail.com

**Keywords:** machine learning, pest introduction, big data, elastic-net

Supplementary material for this article is available online

**OPEN ACCESS**

**RECEIVED**
22 July 2021

**REVISED**
14 September 2021

**ACCEPTED FOR PUBLICATION**
12 October 2021

**PUBLISHED**
29 October 2021



## Abstract

Plant pest invasions cost billions of Euros each year in Europe. Prediction of likely places of pest introduction could greatly help focus efforts on prevention and control and thus reduce societal costs of pest invasions. Here, we test whether generic data-driven risk maps of pest introduction, valid for multiple species and produced by machine learning methods, could supplement the costly species-specific risk analyses currently conducted by governmental agencies. An elastic-net algorithm was trained on a dataset covering 243 invasive species to map risk of new introductions in Europe as a function of climate, soils, water, and anthropogenic factors. Results revealed that the BeNeLux states, Northern Italy, the Northern Balkans, and the United Kingdom, and areas around container ports such as Antwerp, London, Rijeka, and Saint Petersburg were at higher risk of introductions. Our analysis shows that machine learning can produce hotspot maps for pest introductions with a high predictive accuracy, but that systematically collected data on species' presences *and* absences are required to further validate and improve these maps.

## 1. Introduction

Biological invasions describe inadvertent introductions of organisms into new territories. While many entries may not lead to long-term establishment [55], successful establishments of hazardous species can have major consequences for ecosystems and economies [14, 69]. A reliable prioritization of areas for potential introduction would be invaluable to inform surveillance effort [53, 59].

By definition, *introduction* of a species comprises *entry* and *establishment* [23]. Entry of a pest describes its movement into an area and establishment the perpetuation of the species within an area after successful entry [23]. Species distribution models (SDMs)[4] are popular data-driven tools that aim at predicting species' niches on the basis of environmental characteristics of known locations of occurrences [48]. Subsequently, a prediction of the potential area of establishment is derived by assessing the similarity in environmental conditions in other, possibly unsampled, locations. SDMs are commonly developed for specific species. While results from such analyses help to identify risky areas, estimate potential impact and develop management strategies [7, 73], they require species-specific data acquisition, calibration and validation. As a consequence of the time, effort and expertise required for this task, such species-specific analyses are only available for a few hazardous invaders [45]. A generic approach that could help to identify areas that are generally more at risk for pest introduction, without having to first develop a range of species-specific models, would greatly improve evidence-based prevention and management.

The vast majority of SDMs rely exclusively on climatic data to predict where a particular species may establish and maintain a population without the need for further immigration [30]. For invasive species, a growing body of literature stresses the role of anthropogenic factors in the introduction of species [42, 43, 49, 72, 74, 76]. Consequently, such data

[4] Also known as bioclimatic models, climate envelopes, ecological niche models, habitat models, resource selection functions, range maps, among others [18].

could very well improve predictions of hotspots for species introduction [28]. Nevertheless, there have been limited efforts to include anthropogenic features (i.e. predictor variables) into such models [76].

The underlying presence data and the type of features included into the model determines how to interpret results. To predict entry risk, the presence data should only comprise of successful entries of species into non-native territory. The prediction of establishment risk, however, benefits from presence data in native territories to characterize the bioclimatic conditions favored by the species. SDMs based exclusively on climate data map areas' suitability for establishment. While some anthropogenic features are expected to ease establishment, others are related to entry, such as distances to container ports and road density. Consequently, maps derived from a combination of features, where some relate to entry-risk and others to environmental suitability for long-term establishment, depict either entry and/or establishment risk, depending on the feature characteristics in different locations. As an example, areas around container ports might be predicted to be at high risk because many reported presences fall into such areas because of a higher number of successful pest entries, whereas some areas further inland might be predicted to be at high risk because the environmental conditions suit long-term establishment.

Here, we use presence data in both non-native and native territories as well as a large range of features of which some are expected to be related to entry while others are related to establishment. As pest introduction by definition comprises entry and establishment [23], in what follows we refer to introduction-risk to express that our results show both entry-risk and/or establishment-risk depending on the location. Our use of this terminology is therefore in line with the FAO's ISPM definition. Whether higher risk scores are due to entry-risk or establishment-risk is not per se important for our aim of informing surveillance efforts.

We aim to develop a generic modelling approach to identify hotspots for plant pest introductions. We assess the risk of presence in Europe for the whole group of 243 invasive species on the priority lists (A1 and A2) of the European and Mediterranean Plant Protection Organization (EPPO). The A1 list contains species that are absent from Europe while the A2 list contains species with a geographically limited presence. Notably, our objective is not per se to predict the current distribution of these 243 species as well as possible, but rather to use a large set of invasive pest species to derive locations that might be prone to introduction of such species, so-called hotspots, to help predict where future invasions into Europe would be most likely to occur, possibly also for species not included in our data. We obtained worldwide data on the presence of these species from the Global Biodiversity Information Facility (GBIF). Background data[5] were generated using three standard methods recommended by the literature [4, 78]. Global georeferenced data on a wide range of potential predictors related to climate, soils, water, and anthropogenic factors were collected, and an elastic-net machine learning algorithm was trained on around 341 000 observations across the globe to predict new introduction of invasive species as a function of the predictors. The hyperparameters[6] were tuned using three cross-validation techniques. Although the resulting risk maps all have high predictive performance, they show striking differences depending on the background data generating techniques and cross-validation methods considered. Our analysis shows that machine learning can produce hotspot maps for plant pest introduction with a high predictive accuracy, but that systematically collected data on species' presences *and* absences are required to further validate and improve these maps.

## 2. Methods

### 2.1. Data

#### *2.1.1. Species presence*

The list of species was obtained from the A1 and A2 list of EPPO (version 2020-09)[7]. Both lists contain species that are recommended for regulation as quarantine pests in Europe. The A1 list contains species that are absent from Europe while the A2 list contains species with limited presence. Subsequently, on the 30th of March 2021, 490 323 worldwide occurrences of these species were obtained from the Global Biodiversity Facility (GBIF) (10.15468/dl.fc5kva). The raw data was cleaned by removing all points with any of the following characteristics: reporting year prior to 1970, fossil specimen, literature-based observations, preserved specimen, location falling exactly on the centroids of capitals, or centroids of countries, or into sea, or on biodiversity institutions assuming that those are part of a collection [84]. Furthermore, presences with duplicated values across all features (i.e. input variables of the models) were removed,

[5] Background data characterize the feature space and act as pseudo-absences to which presence data are compared within the classification model. They do not necessarily aim to be true absence points, but rather provide a characterization of possible values features could take throughout the studied geographic area.

[6] The term hyperparameter denotes a parameter that controls the learning process of the algorithm but that is not directly inferred from the training (i.e. fitting) of the model as is the case for coefficients. In other words, hyperparameters hold settings that influence the structure of the model. A standard approach is to tune these hyperparameters (i.e. optimize) by running the learning algorithm for different values and choosing the hyperparameter value that results in the best performance according to a cross-validation procedure. The elastic-net has two hyperparameters (section 3.4).

[7] www.eppo.int/ACTIVITIES/plant_quarantine/A1_list; www.eppo.int/ACTIVITIES/plant_quarantine/A2_list.

thereby, we effectively thinned presences at the scale of the finest environmental predictors. Removing duplicated datapoints is considered good practice in machine learning applications because duplicated entries hold little information while potentially biasing the prediction and inflating performance. Furthermore, as GBIF is a collection of various datasets removing duplicated datapoints eliminates the risk that the same individuals were observed and reported by several people, over multiple years, and/or in different datasets. In addition, the spatial nature of the GBIF database results in autocorrelated presences. This autocorrelation is usually reduced by thinning the presence points before computing SDMs. Using the spatial resolution of the feature layers is often used to determine the thinning radius [73]. The presence thinning can be achieved in a computationally faster way by omitting duplicated datapoints. Lastly, all presences from 2020 and 2021 were removed and presences in Europe for these two years were used for testing model performance (see section 2.5). The final set of 170 460 presence data, for which complete and unique combinations of feature data were available, spans 243 species, 92 families, 52 orders, 21 classes and 13 phyla, or more specifically 133 *Arthropoda*, 37 *Tracheophyta*, 2 *Mollusca*, 18 *Ascomycota*, 1 *Negarnaviricota*, 15 *Basidiomycota*, 19 *Proteobacteria*, 4 *Oomycota*, 2 *Cressdnaviricota*, 5 *Nematoda*, 2 *Actinobacteriota*, 4 *Kitrinoviricota*, and 1 *Chytridiomycota*. For model training, we classified presence of any species as a 1 and pseudo-absence as 0 (see next section).

#### *2.1.2. Background data*

The GBIF data usually[8] come as presence-only and it was thus necessary to generate background data representing pseudo-absences to train and test our models, as commonly done in SDMs [4]. While this is common practice in the SDM literature, there is no consensus regarding the best approach [70]. The issue of generating background data is particularly difficult to resolve when presence data is biased due to spatial variation in reporting [78]. While GBIF is extensively used in ecological research [32], geographic bias is very likely [8, 24]. We tested three ways to generate pseudo-absence data all of which find support in the literature [4, 68, 78].

First, random data were generated on a global scale covering all parts of the world except the poles (supplementary material: figure S1 (available online at stacks.iop.org/ERL/16/114026/mmedia)). Randomly sampling background data (denoted random) is the default strategy in SDMs and frequently recommended (e.g. [4]). The approach implicitly assumes that the entire geographic extent is equally relevant for the analysis and that the entire possible feature space should be used as a comparison to the features of the presence locations. Depending on the data generating process of the presence-data, this assumption might not be appropriate. Often, presence-data are not collected following a strict sampling protocol. In opportunistic sampling, people visit some places more than others, for example due to ease of access or aesthetical reasons. This geographic bias results in an environmental bias that can result in biased predictions [11].

Second, conceptually close to the bias-file approach of the popular SDM algorithm MaxEnt, we generated data from a biased background which aims at mimicking the geographic bias in the GBIF database [68, 78]. Here, presences were counted within 5 decimal degree grids. Next, a two-dimensional Gaussian kernel density was estimated on the count-grids and rescaled such that all values sum to unity. Subsequently, background data points were generated by sampling from this estimated spatial probability distribution. With this second approach (denoted kdbias), the background data tend to remain close to the presence data, as would be the case if the sampling areas were kept close to each other (supplementary material: figure S2). Thereby, an implicit assumption is made that only areas nearby known presences are relevant for the analysis and that the feature space used as a comparison to the presences should be restricted to nearby conditions. By sampling the pseudo-absences from a background that mimics the sampling effort of the presence-data, predictions may become unbiased [11]. However, this technique to pseudo-absence generation will always result in many points that are known to be false negatives. As data generated in this approach will have many locations with, both, presences and pseudo-absences (supplementary materials: figure S2), performance metrics will be over-pessimistic.

Lastly, we combined the biasing approach with Barbet-Massin *et al*'s [4] recommendation for geographic exclusion (denoted kd05dfar). Here, we generated a larger number of data from a biased background and subsequently removed data that were less than 5 decimal degrees away from any presence data (supplementary material: figure S3). From the remaining background data, a random subset was sampled such that the resulting data had a balanced number of presences and background data. Notably, while Barbet-Massin *et al* [4] recommended a distance of two degrees, in latitude or longitude, this criterion would have resulted in a questionable comparison in our case due to the large number and geographic spread of species presences in our database. Nevertheless, our approach intends to provide more background data within proximity of the presences [78], without heavily overlapping background data with presence data as done in the pure biasing approach as employed in MaxEnt. Hence, as with the

[8] Our data comprised 264 absences which were removed to ensure methodological consistency across all the pseudo-absences.

second approach, the implicit assumption is made that areas nearby known presences are more relevant for the analysis, thereby, addressing the spatial heterogeneity of opportunistically sampled presences [11]. However, here the feature space used as a comparison to the presences does not comprise conditions of areas where pest presence is reported which should provide less pessimistic measures of performance.

The final datasets had a balanced distribution of presences and pseudo-absences, i.e. a sampling prevalence of 50% [4, 60].

*2.1.3. Features*

Various georeferenced data were gathered. Table S1 in the supplementary material provides an overview for the features and table S2 for the raw data. Data on climate were obtained from Karger *et al* [47]. Soil characteristics were obtained through OpenLandMap [1, 33, 34, 36–41, 83]. An indicator of erosion risk was obtained from the World Resource Institute [82]. Information on landcover was obtained from Buchhorn *et al* [12]. A dataset on water related indicators was obtained from the World Resource Institute [81]. An indicator of biodiversity intactness was obtained from Newbold *et al* [63]. Data on population density were obtained from the Joint Research Centre [22]. Data on road densities for different road types were obtained from Meijer *et al* [61]. An indicator of anthropogenic pressure on the environment was obtained from Venter *et al* [77]. Data on human-driven modification of terrestrial systems was obtained from Kennedy *et al* [50]. A spatial layer on accessibility to cities, measured in driving time, was obtained from Weiss *et al* [80]. Studies advocated for the use of the gross domestic product (GDP) in analyses of invasive species [27, 45]. However, GDP is generally only available at course, country-level, resolution. Therefore, we decided to proxy GDP using spatial data on radiance of nightlights which were obtained from Hengl [35]. Various studies have found high correlations between the nightlight radiance and the GDP and expressed support of using this feature as a spatially explicit proxy for GDP [9, 10, 25]. Lastly, georeferenced data for container ports was obtained from Bartholdi *et al* [5]. These data comprise longitude and latitude as well as connectivity indices for 200 container ports around the world. For each presence and background point, the minimum distances to a port and the mean distance to all ports were computed. Subsequently, connectivity indices of the closest port, as well as connectivity indices for all ports, weighted by their inverse distances to a particular point, were used as features.

### 2.2. Data processing

First, observations with incomplete data were omitted. For categorical features, the 19 most frequent categories were kept, and other categories were grouped into one. Next, categorical features were dummy encoded. In addition to the spatially weighted port connectivity indices, the following continuous features were engineered: the average annual photosynthetically active radiation, the standard deviation of the photosynthetically active radiation across months, the change in population density between 1975 and 2015, and the change in human impact on the environment between 1993 and 2009. All continuous features were transformed to normality, centered, and scaled. The best transformation to normality was estimated from a set of candidate functions using only the training data [66]. The final datasets, for the three approaches to background data, have 340 920 points, half of those being presences and the other half background data, with complete data for 246 features. Out of those, 181 features were continuous, and 65 features were dummy encoded categories.

### 2.3. Cross-validation techniques

We implemented and compared three cross-validation techniques to optimize the model hyperparameters (i.e. the parameters of the penalty term of the elastic-net). First, we followed the most widely used approach of randomly splitting the data into folds[9] [71]. To manage computational time, we used five folds. Second, we separated data into continental spatial blocks (supplementary material: figure S10). Here, six folds were generated corresponding to the continents Africa, Asia, Australia, Europe, North America, and South America. As such, we intended to assess the transferability of the model across geographic space [62, 67, 71]. Lastly, we used temporal splits for cross-validation in which presences were separated by their year of record and background data randomly assigned, without replacement, such that balanced folds were obtained. Due to the exponential increase in presence records over time (supplementary material: figure S4), we divided them into unequal time periods corresponding to the years 1970–2005, 2006–2011, 2012–2014, 2015–2016, 2017–2018, and 2019, resulting in six folds with an approximately equal number of presences in each. Through forward chaining of the temporal folds[10], we intended to test

[9] A fold is a term used in machine learning to describe subsets of the data. For example, five randomly split folds correspond to five data partitions each holding 20% of the training data.

[10] We refer to *forward chaining* to describe an out-of-sample approach in which the temporal order of the cross-validation folds is considered. In the first iteration, the cross-validation starts by training the algorithm on data from the first time-period and validating performance on data from the second time-period. In the second iteration, data from the first and the second time-period is used for training and performance validated on data from the third time-period, and so on. Hence, data available for training is growing over time. The first time-period is not used for validation, whereas the last time-period is not used for training within the cross-validation. This cross-validation, and comparable variations, is commonly used for time-series analyses. The technique is reviewed under the name *prequential growing window* within Cerqueira *et al* [13].

the model's ability to predict future introductions (supplementary material: figures S5–S9).

### 2.4. Algorithm and hyperparameter tuning

The model is a generalized linear model based on a logit link, equivalent to a logistic model. The model includes regression coefficients that are estimated using a learning algorithm called elastic-net [26, 85]. The algorithm is a regularization technique that combines the L1 (sum of absolute coefficient magnitudes) and L2 (sum of squared coefficient magnitudes) coefficient penalties into the loss function. In doing so, the model is a generalization of the lasso and ridge regression approaches and allows for the estimation of pure versions of the two as well as mixed variants.

We decided to use this training algorithm because of its ability to find an optimal balance between bias and variance. The elastic net reduces variance at the cost of introducing bias to minimize the prediction error. This approach is called *regularization* and is designed to optimize the predictive performance of the model. The algorithm is computationally relatively fast, memory efficient, and robust to correlated features. It is thus well-adapted to large scale practical applications.

The parameters are estimated using a penalized log likelihood objective function [85]. The likelihood is based on a binomial distribution and the penalization is based on the elastic net penalty. The elastic net includes a penalty term defined by two hyperparameters named $\alpha$ and $\lambda$. The hyperparameter $\alpha$ describes the mixing of the L1 and L2 penalties. If $\alpha$ equals 1, the elastic net would essentially be a lasso regression whereas $\alpha$ equal to 0 would result in a pure ridge regression. The hyperparameter $\lambda$ denotes the degree of regularization employed. In the elastic-net algorithm, the regularization determines the extent to which coefficient magnitudes affect the loss function. Consequently, the regularization determines the extent to which coefficients are shrunk toward zero. By shrinking coefficient values, a model fit is obtained that might generalize the underlying relationships better.

Both hyperparameters were tuned using a grid search to maximize the AUC value computed successively with the three above-mentioned cross-validation techniques. In principle, the AUC metric measures the correctness in rankings between locations which is directly related to our modelling objective of identifying areas at risk [2, 68]. However, whenever true absences are not available, performance metrics represent heuristic measures only and should therefore be cautiously interpreted [54, 62]. This is because classification-based performance metrics such as AUC are not only based on the correct classification of the presence-class but also the absence-class. Without true absences, however, the true number of misclassified absences remains unknown. Sensitivity and specificity receive equal attention in our results. The presented values correspond to values obtained at a cut-off of 0.5. For $\alpha$, values between 0 and 1 were searched at 0.1 increments and for $\lambda$ values between 0 and 1 at 0.025 increments, resulting in a total of 451 combinations. In the supplementary material, figures S14–S22 depict the tuning results, and table S3 depicts the optima, for each cross-validation technique and background data generation approach. Only the spatial block cross-validation resulted in regularized models, while the random and temporal splits suggested that no regularization yielded the best performance. No regularization ($\lambda = 0$) essentially collapses the elastic net into a standard generalized linear model with binomial distribution. The tuning results for random and temporal cross-validation could be related to the spatial clustering of data which resulted in small regularization values in other studies [2]. While the lack of regularization was less surprising for the random cross-validation, it did surprise us for the temporal cross-validation splits. The tuning-results for the temporal cross-validation essentially suggest that the spatial patterns of invasive species reporting were so stable over time that the algorithms did not need generalization capability by limiting the model flexibility (i.e. regularization) to extrapolate to future time periods. Following the hyperparameter tuning, the model coefficients were estimated using the entire training data.

### 2.5. Test data

It is common practice in machine learning applications to test the algorithms on data which were entirely withheld during the processing, tuning, and training steps. These data are referred to as test data. The classification performance on the test data intends to provide objective measures of model reliability for the task at hand. Consequently, the test data must be constructed in ways that correspond to the intended use of the model. Ideally, test data should be completely independent of the training dataset. Unfortunately, in many cases this is not possible due to a lack of data availability, and practitioners often rely on splitting part of the dataset intended for training, withholding the split data during processing, tuning, and training steps, and using these data only for testing the final performance.

As mentioned above, our modelling objective is the prediction of risk of invasive pest introduction across Europe. Hence, the risk map could support surveillance if predictions were to describes introduction risk in future time periods well. Therefore, the ideal test data would comprise systematically sampled records which hold novel invasive pest introductions in Europe sampled in recent months which were not included in the timeframe of the training dataset. We were unable to obtain results of systematic invasive species surveys from the National Plant Protection Agencies. Furthermore, introductions of new

invasive pest species (fortunately) remain rare events. Consequently, generating a test dataset with a sufficiently large number of locations which correspond to recent introductions of invasive pests is expected to be difficult even if pest survey data across Europe would be accessible.

Considering the unavailability of systematically sampled and fully independent test data, we decided to omit all data for the years 2020 and 2021. Subsequently, we used presences in 2020 and 2021 which fall into Europe to test the final models' performance. A random subset of the pseudo-absences was assigned to the test data such that a balanced dataset was obtained. In doing so, we generated a test dataset with 13 158 European points with half being presences and half pseudo-absences. Notably, majority of the European presence data in 2020 and 2021 does not represent reporting of newly introduced invasive pest species but merely new records of invaders already sighted in earlier years, albeit at different locations. Hence, it is possible that temporal dependencies between training and testing data resulted in too optimistic test performance.

### 2.6. Prediction and mapping

To circumvent the problem of differences in the resolution of input layers, longitude, and latitude coordinates for around 870 000 points across Europe were generated. The number of points was chosen such that the modelling steps are feasible in terms of computational time and memory requirements. Subsequently, for all points feature data were extracted and processed as described above. To minimize empty spaces in the risk map, due to systematically missing input data in certain locations, individual features were imputed for 73 380 points, with values of the geographically closest point within a maximum distance of 1 decimal degree. Points with partially missing data mostly fall on coastlines and on in-land waterbodies. Consequently, missing data is likely due to resolution-related artifacts of pixels which fall on non-linear country borders and unavailable information for some features.

The trained and tuned models were used to generate a continuous probability score for introduction at all points in Europe. All maps are point-based. Each point was coloured using the probability score. The figures shown within the manuscript depict the average probability score across the three background data approaches, for models tuned on temporal and continental cross-validation techniques. The sensitivity of this probability to the background data approach is shown through visualizations of the range of the probability score which was computed by taking the difference between the maximum and minimum values for each point. Individual maps for all approaches are provided in the supplementary material (figures S23–S37).

## 3. Results and discussion

### 3.1. European hotspots for pest introductions

As our objective is the analysis of hotspots to improve the management of future introductions, we believe that temporal cross-validation most closely represents our objective. However, the spatial-block design best mimics spatial transferability[11] [19, 70, 71]. We will discuss average results, across different background generation approaches, for models tuned on temporal and continental splits.

Figure 1 depicts the average predicted risk of new introduction, across the three approaches for background data generation, for models tuned using temporal and continental cross-validation. Irregular, polygon-like, surfaces in the maps result from input data on water indicators which came in the form of spatial-polygons. More importantly, hotspots, i.e. areas with high probability of presence of at least one invasive species, were consistently predicted to fall into highly anthropogenically-impacted areas. The BeNeLux states, Northern Italy, the Northern Balkans, and the United Kingdom were generally predicted to be at higher risk of future introductions. The contrast between regions at low and high risk was higher in models tuned on temporal folds compared to models tuned on spatial folds.

### 3.2. Feature contributions

The importance of each feature (i.e. variable) was computed using the Feature Importance Ranking Measure [29]. Here, we discuss feature contributions based on their average score across the different background generation approaches.

For models tuned on temporal splits, the highest ranked features were temperature-related features and soil sand content at one meter depth. Locations characterized by sandier soil were associated with higher risk scores. Many of the analysed species are forestry pests. As forests are often characterized by sandier soils, it could explain why higher values for the soils' sand content were found to increase risk. Minimum, average, and maximum temperatures in the different months had varying effects. For example, higher minima in February consistently increased risk while higher minima in January consistently decreased it. Next to temperature-related features and soil sand content, port connectivity, water availability, water withdrawal, soil water content, access to cities, the minimum distance to a port, a spatial proxy of the gross domestic product (GDP), and the road

[11] Transferability describes the ability of the model to generalize and correctly predict new areas or time periods.

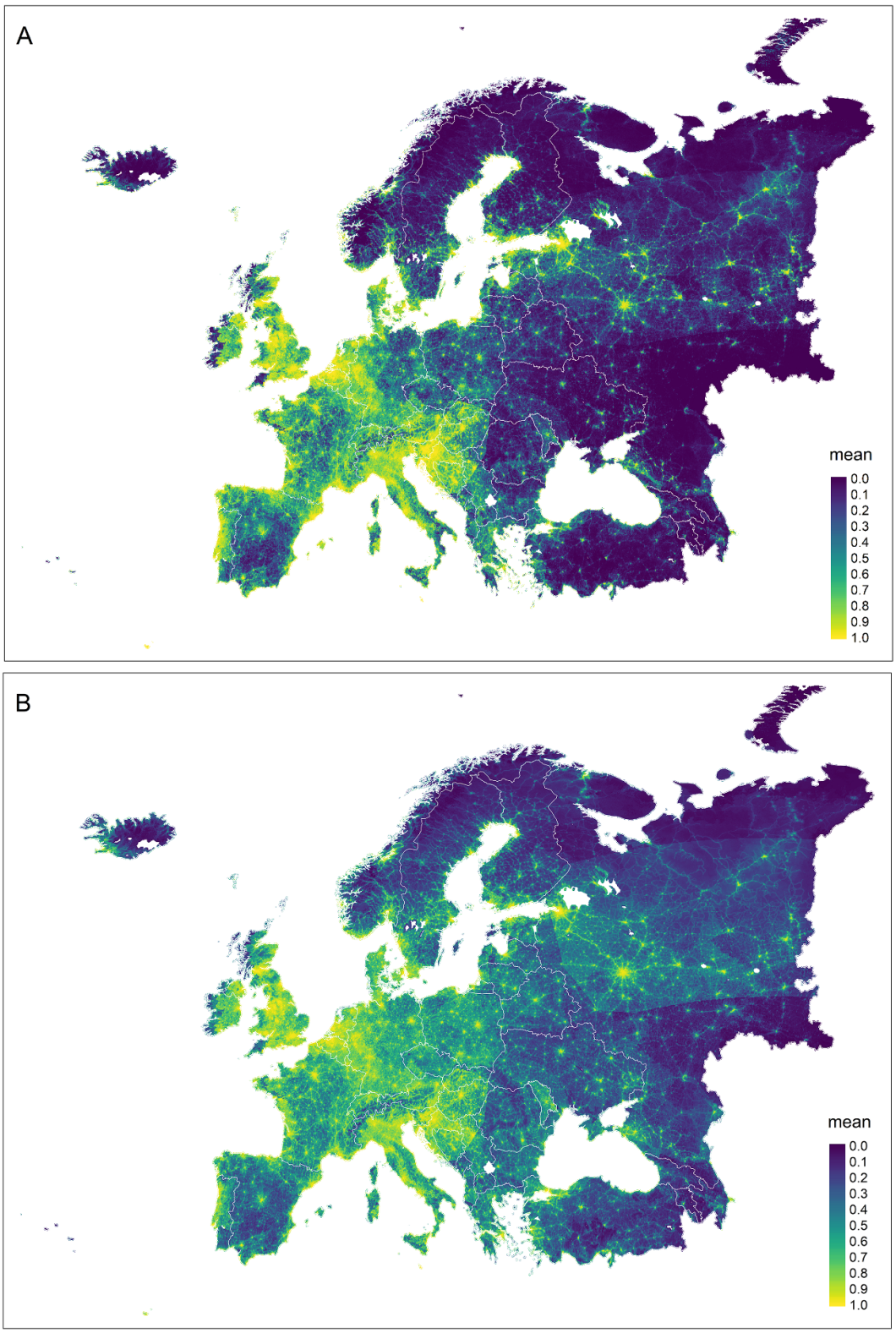


**Figure 1.** Probability of new introduction across Europe. The mean risk levels across different models tuned using temporal (a) and spatial (b) cross-validation.

density, among others, were important. Higher values for anthropogenic features were generally associated with higher risk scores[12].

For models tuned on continental splits, the feature ranking differed considerably compared to the models tuned to predict into future time periods.

[12] Access to cities and minimum distance to a port are inversely related to anthropogenic pressure as higher values correspond to longer driving times to a city and larger distances to a port, respectively.

**Table 1.** Overview of model performances for all cross-validation and background generation techniques. Performance was measured by the area under the ROC curve computed by cross-validation or using an independent test dataset. An AUC of 1 indicates perfect classification while an AUC of 0.5 indicates random classification.

| | | Cross validation | | | Test data[a] | | |
|---|---|---|---|---|---|---|---|
| Cross validation | Background data | AUC | Sens. | Spec. | AUC | Sens. | Spec. |
| Random | Random | 0.98 | 0.95 | 0.93 | 0.93 | 0.93 | 0.93 |
| Random | kdbias | 0.91 | 0.83 | 0.84 | 0.76 | 0.75 | 0.76 |
| Random | kd05dfar | 0.99 | 0.95 | 0.96 | 0.89 | 0.85 | 0.93 |
| Spatial | Random | 0.95 | 0.77 | 0.91 | 0.93 | 0.97 | 0.89 |
| Spatial | kdbias | 0.84 | 0.65 | 0.83 | 0.74 | 0.79 | 0.69 |
| Spatial | kd05dfar | 0.95 | 0.77 | 0.88 | 0.86 | 0.93 | 0.80 |
| Temporal | Random | 0.97 | 0.88 | 0.94 | 0.93 | 0.93 | 0.93 |
| Temporal | kdbias | 0.87 | 0.70 | 0.86 | 0.76 | 0.75 | 0.76 |
| Temporal | kd05dfar | 0.98 | 0.89 | 0.96 | 0.89 | 0.85 | 0.93 |

Sensitivity (Sens.) and specificity (Spec.) were computed for a threshold of 0.5.

[a] Test data refers to European data in 2020 and 2021.

Here, anthropogenic features dominated the ranking. Across all approaches to background data, the degree of nightlight radiance, being our spatial proxy of GDP, ranked very high as a risk increasing factor. Accordingly, access to cities, minimum distance to a port, road densities for various road types, water withdrawal, the human impact on the environment, and the population density were important features. In general, effect directions again suggested that areas with a higher anthropogenic impact are at a higher risk. Next to anthropogenic features, higher values for drought severity, seasonal water variability, elevation, and the land cover classification for moss and lichen as well as cultivated and managed cropland decreased risk, while the biome classification for temperate sclerophyll woodland and shrubland, and higher values for flood occurrences, biodiversity intactness, organic carbon content in the soil, average photosynthetically active radiation in September, and soil water content were associated with increased risk.

### 3.3. Model performance

Model performance depended on both the cross-validation technique and the background generation approach (see table 1). In terms of cross-validation, the highest performance scores were obtained by randomly splitting data into folds, followed by temporal splits and lastly continental splits. Randomly splitting not only resulted in higher average performance scores across validation folds, but also in a severely reduced variation of performance across folds compared to the temporal and continental techniques (supplementary material: figures S11–S13). The high performance with random splitting is likely related to spatial clustering of species presence. This violates the independence assumption and leads to models that overfit to residual dependencies, resulting in overoptimistic model performance [71]. Within the continental cross-validation, the validation-scores obtained for the European continent were AUCs of 0.97, 0.76, and 0.93 for the random, kdbias, and kd05dfar background data, respectively. In other words, the algorithms predicted European data well after being trained exclusively on data in other parts of the world.

Concerning the background generation approach, the highest performance scores were obtained with the random approach, followed by the geographic exclusion approach (kd05dfar), and lastly the biased data generation technique (kdbias) (see table 1). While the very good performance of the random technique is likely inflated by the large geographic scale considered here [6, 75], the lower performance for the biased approach is arguably over-pessimistic as the approach results in a large number of data points in the exact same locations yet opposing classifications for the dependent variable (supplementary material: figure S2).

While the temporal and spatial block approaches did result in lower cross-validation performance scores compared to randomly split folds, they test and optimize traits of the model that are desirable for our purpose which led us to present the models above. The performance of models to predict introduction into new geographic spaces or to provide a prioritization of areas for future introductions, is most appropriately estimated by cross-validation techniques that also simulate those behaviours. In addition, cross-validation techniques that simulate the modelling objective result in hyperparameter values that are optimized for the task. As a result of the hyperparameter values, feature selection and model fitting are optimized for the research objective as well. Having said that, as mentioned above, the temporal patterns of invasive pest reporting were so stable over the years 1970–2019 that the temporal cross-validation, like the random cross-validation, did not lead to regularized models. Because of the same tuning results, predictions of the temporal and random

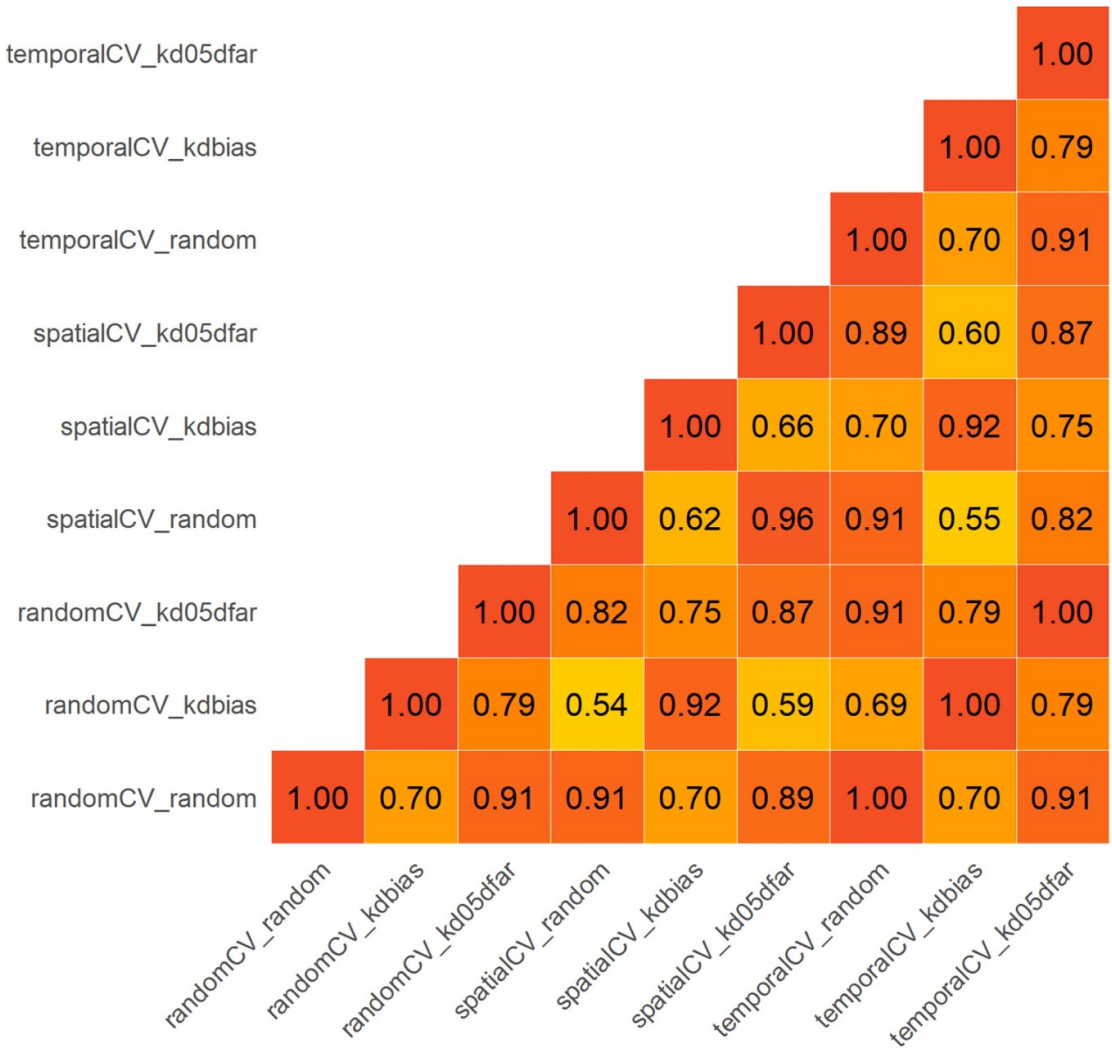


**Figure 2.** Spearman correlation of the European prediction across all combinations for cross-validation and background data generation.

cross-validation were the same which resulted in the same test performance as well. While for the data used here the two cross-validation techniques did not produce different predictions, this could very well be different in datasets which require more generalization capacity across time-periods. Therefore, despite the equivalent results obtained here for the random and temporal cross-validation, we urge practitioners to align their cross-validation approach with their modelling objective.

Performance was reasonably good, even for rigorous validation approaches such as temporal and continental splits, indicating that top-down analyses, through the bundling of species, do not necessarily sacrifice performance per se. Interestingly, global presence patterns were quite stable over time (supplementary material: figures S5–S9). As hotspots for pest introductions did not change considerably over the time horizon 1970–2019, the cross-validation scores obtained from temporal splits suggest that the models are very much able to predict future introductions based on past ones.

### 3.4. Sensitivity

Despite their somewhat comparable accuracies, the generated risk maps as well as the importance of the features of the corresponding models were drastically different (figure 2). Similar to Austin [3], our analysis shows that equivalent performance metrics can result in very different models and outputs.

Figure 3 depicts the difference between the maximum and minimum probability values across the three approaches of generating background data for models tuned using temporal and continental cross-validation. The different background generation approaches resulted in sizable changes in predicted risk for large parts of France, Germany, Northern Spain, and Moldova. Individual maps for all approaches are provided in the supplementary material (figures S23–S37).

Figures 4 and 5 depict the 50 most important features, on average across different approaches to generating background data, for models tuned using temporal and continental cross-validation, respectively. The importance of features, and occasionally

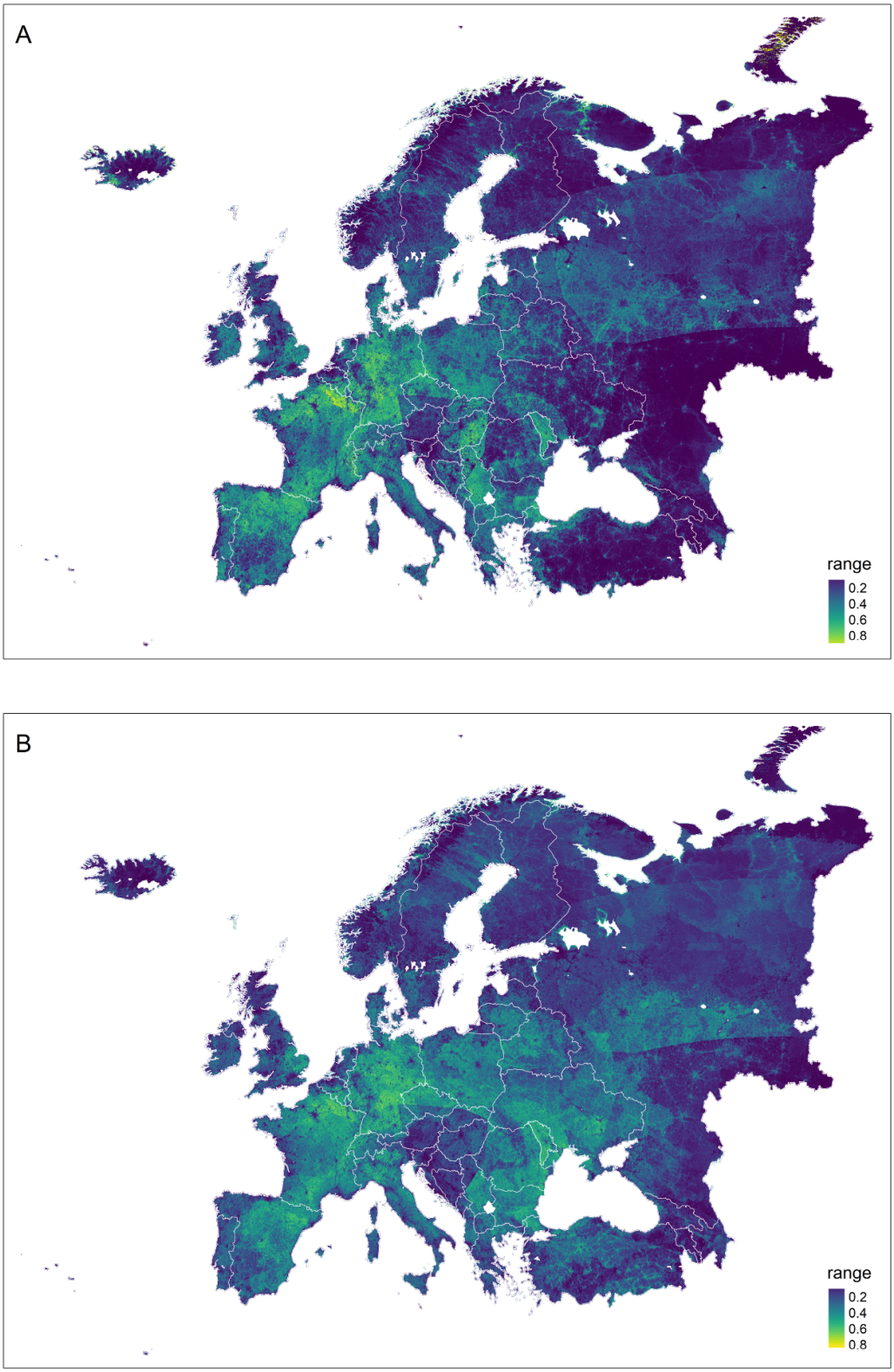


**Figure 3.** Difference between maximum and minimum risk level values for Europe across different background generation approaches for models tuned using temporal (a) and continental spatial-blocks (b) for cross-validation.

coefficient directions, varied considerably suggesting that very different models were created. See figures S39–S45 in the supplementary material for further examples of feature importance in different models. The results stress the diversity in models that can be built using the same presence data. Arguably, this underscores the need to explore sensitivity of results beyond computing several learning algorithms using the same data generating process, especially if clear data on true species absence are unavailable.

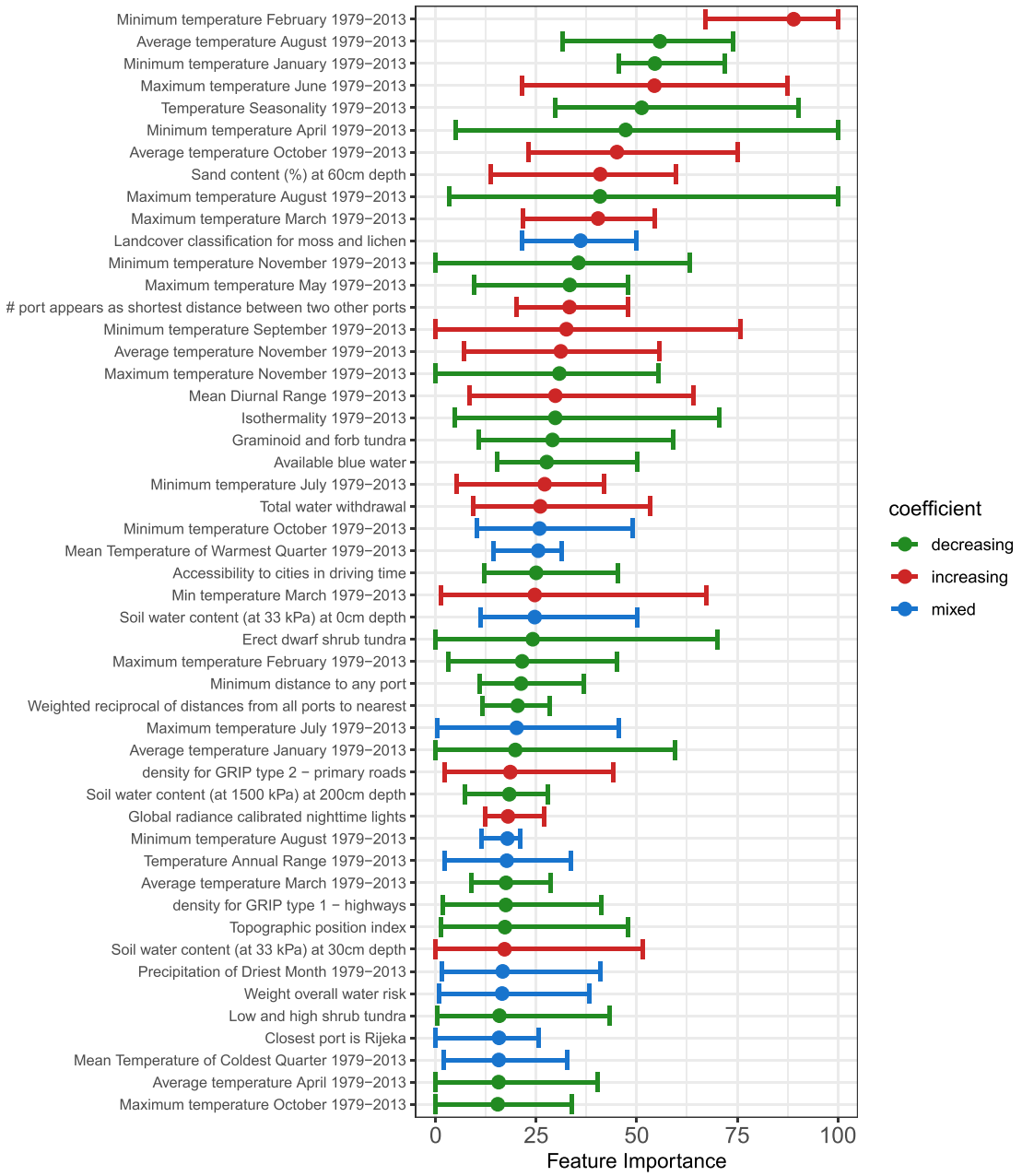


**Figure 4.** Fifty most important features across different background generation approaches for models tuned using temporal split cross-validations. The average is depicted as a dot. The line range shows the minimum and maximum values. If coefficients were consistently negative (i.e. risk decreasing) across all approaches for background data generation, the graph is coloured green. If coefficients were consistently positive (i.e. risk increasing) across all approaches, the graph is coloured red. If the coefficient direction changed between the different approaches to generating background data, it is coloured blue.

### 3.5. Implications for pest surveys

The lack of systematically surveyed species presence and true absence restricted us from disentangling whether predictions were a result of monitoring or reporting bias, or of area characteristics that indeed promote the introduction of invasive species. As an example, areas around container ports such as Antwerp, London, Rijeka, and Saint Petersburg, were generally predicted to be at high risk. The literature frequently discusses the involvement of international trade [42, 43, 65], in particular via boats and roads [27, 49], in the introduction of invasive species. Ecosystems characterized by a high level of anthropogenic disturbance are expected to facilitate species entry and establishment [49, 65]. Consequently, our results would be in line with the expectations from the literature. However, because regulators, scholars, and citizens expect that these areas likely contain new introductions, these locations are also often highly monitored, which could lead to reporting bias (supplementary material: figure S38).

Systematic survey data on species presences and, equally important [56, 76], true absences would suspend these concerns entirely. Such data would allow to measure to what extent pest presences are driven by anthropogenic features, without having to ponder

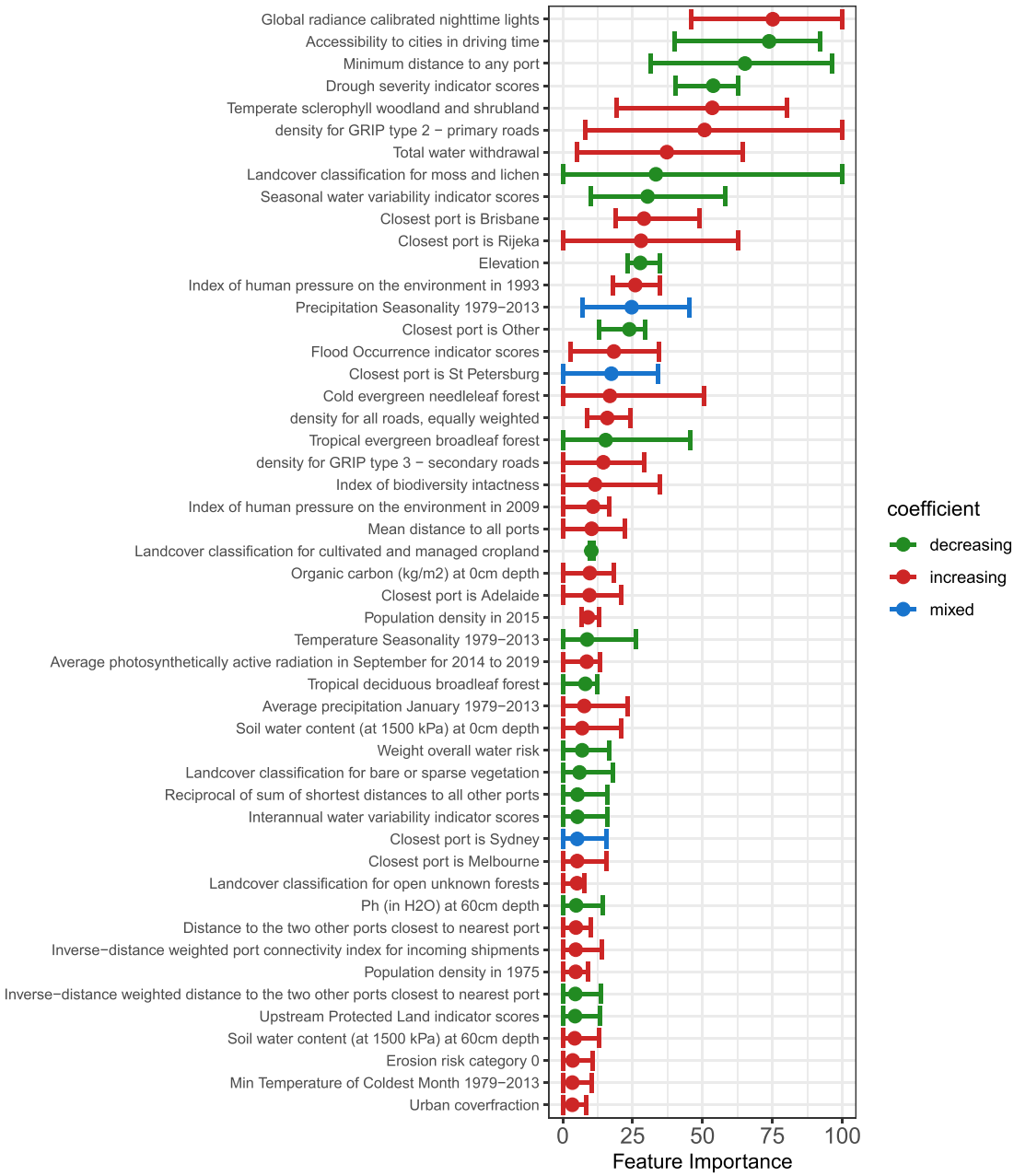


**Figure 5.** Fifty most important features across different background generation approaches for models tuned using spatial-block split cross-validations. The average is depicted as a dot. The line range shows the minimum and maximum values. If coefficients were consistently negative (i.e. risk decreasing) across all approaches for background data generation, the graph is coloured green. If coefficients were consistently positive (i.e. risk increasing) across all approaches, the graph is coloured red. If the coefficient direction changed between the different approaches to generating background data, it is coloured blue.

whether these characteristics were exclusively, or partially, related to biases [19, 20, 24, 79]. The inclusion and analysis of anthropogenic features is critical to further our understanding of externalities from human-driven land-use change, infrastructure, and trade. Efforts to include anthropogenic features into models, except for attempts to correct for data biases, are lacking [76]. The unavailability of systematic data for the left-hand side of the equation is likely a major reason for that.

Absence of a species may be due to one of the following causes [56]. First, *environmental absence* describes locations with unsuitable environmental conditions. Second, *contingent absence* describes locations which are suitable per se, but due to dispersal limitations, local extinctions, or an inadequate size of the suitable patch, among other factors, they remained free of the species at the time of observation. Lastly, *methodological absence* describes locations which are falsely classified due to underlying biases, or incomplete coverage, in the available calibration data. SDMs predicting the fundamental niche aim to correctly classify environmental absences from presences, whereas contingent absences become particularly important when predicting the realized niche [56, 76]. Methodological absences taint

predictions regardless of modelling purpose yet are likely to prevail in most data used for SDM research [46]. While appropriate surveying for true absences requires considerable labour input [58], without such data, predictions may be biased by the spatial variability in opportunistic sampling and, as a consequence of the unknown false negative rate, true model performance remains unknown.

Global estimates suggest that the impact of invasive species runs in the trillions of Dollars [17]. For Europe, conservative estimates of annual impacts range from 12.5 to 20 billion Euro [21]. Several thousand species have already invaded Europe and the annual rates of new establishments are progressively increasing [45, 49]. The continuous rise in flow of products and people will likely only aggravate the risk of biological invasions in the future [15, 16, 43, 44]. Nevertheless, compared to estimates on current and future impacts, expenses for management and surveillance remain low [51]. While the process of hazardous invasions will remain random, predictive models in combination with the ever-increasing amount of georeferenced data can improve support of decision making in the future.

Harmonizing species surveys and making the resulting data available for research can further improve the prediction of hotspots. For invasive species on EPPO's priority lists, annual surveys are already conducted by the European member states. These data remain unharmonized across states, inaccessible to researchers, and without records of true absences. The inclusion of true absences in such efforts is as important for predictive models as the systematic collection of presences [56, 76].

This study shows that machine learning methods allow for the generation of generic risk maps for invasive pest introduction that represent the underlying data with high predictive accuracy into the future or into new territories. However, the usefulness of such maps in practice depends critically on the quality of the underlying data. Ground-validation through presence/absence data from systematically sampled observations in the field is required to enable sound judgement and decision making based on model predictions. Results of annual pest surveys could iteratively be used to test the current models and subsequently feed into updating the risk map. In doing so, the risk map could dynamically aid in the continuous surveillance of hotspots of invasive pest introductions and help gain a better understanding of which characteristics of these locations lead to more invasion events. Hence, data-driven machine-learning predictions and ground-validation are not substitutes but complements in the goal of understanding and preventing invasive pest introductions.

While our analysis is a critical call for the need of systematic survey data, we believe the obtained results are a reason for optimism. In the last decades, previously unimaginable advances have been made in the breadth and quality of georeferenced environmental and anthropogenic data and computing technologies. Consequently, the quality of our predictions is more than ever bottlenecked by the lack of open data on results of systematic surveys and records on absence. Considering the current and potential future impact of invasive species to our ecosystems and economies, additional funding for species surveys would likely result in significant paybacks by informing the design of management strategies using predictive models.

Next to the data-needs described above, there are several important avenues for future work. First, the value of additional information on areas' general risk of pest introduction is largely determined by the actions regulators and risk managers take based upon this new information [76]. Consequently, future research should investigate in which ways data must be communicated and integrated into the risk managers' workflow to enable improved decision-making. Second, hotspots of pest introductions may shift under a changing climate. Therefore, future work could project potential changes in pest introduction hotspots for different climate change scenarios to support preparedness. Similarly, shifts in trade patterns, or human behaviour, could be analysed if future projections of anthropogenic features were available. Lastly, data collection through citizen science holds great potential because such efforts can scale immensely [52]. However, for such data potential reporting biases must be addressed ideally at the stage of the data generating process. Future research on technological approaches, such as mobile applications [31, 57, 64], which might aim at alleviating opportunistic sampling by coordinating citizens' search efforts into a more systematic spatial and environmental coverage would be invaluable.

## 4. Conclusion

Pest risk assessments are commonly developed for individual species. An overwhelming absence of information on areas risk toward invasive species introduction results out of the significant time and labour requirements of species-specific analyses which complicates management. We aimed to develop a generic modelling approach to identify hotspots for plant pest introductions. We assessed the risk of presence in Europe for the whole group of 248 invasive species on the priority lists (A1 and A2) of the European and Mediterranean Plant Protection Organization. Global georeferenced data on a wide range of potential predictors related to climate, soils, water, and anthropogenic factors were collected, and an elastic-net machine learning algorithm was trained on around 341 000 observations across the globe to predict new introduction of invasive species

as a function of the predictors. The algorithm was tuned and trained for nine setups resulting from the combinations of three approaches to generating background data and three cross-validation techniques.

Results revealed that the BeNeLux states, Northern Italy, the Northern Balkans, and the United Kingdom, and areas around container ports such as Antwerp, London, Rijeka, and Saint Petersburg were at higher risk for introductions. For models tuned to predict into future periods, the highest ranked features were related to temperature. For models tuned to predict into other continents, anthropogenic features such as the degree of nightlight radiance, access to cities, minimum distance to a container port, road densities, the human impact on the environment, etc dominated the feature importance ranking.

Harmonizing species surveys and making the resulting data available for research can further improve the prediction of hotspots. For invasive species on European and Mediterranean Plant Protection Organization's priority lists, annual surveys are already conducted by the European member states. These data remain unharmonized across states, inaccessible to researchers, and without records of true absences. Our analysis shows that machine learning can produce hotspot maps for pest introductions with a high predictive accuracy, but that systematically collected data on species' presences *and* absences are required to further validate and improve these maps.

## Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.5554971. Data will be available from 01 October 2021.

## ORCID iDs

Kevin Schneider https://orcid.org/0000-0001-7170-218X
David Makowski https://orcid.org/0000-0001-6385-3703
Wopke van der Werf https://orcid.org/0000-0002-5506-4699

## References

[1] Amatulli G, McInerney D, Sethi T, Strobl P and Domisch S 2019 Geomorpho90m-global high-resolution geomorphometry layers: empirical evaluation and accuracy assessment *Technical Report PeerJ Preprint*
[2] Anderson R P and Gonzalez I 2011 Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent *Ecol. Modelling* **222** 2796–811
[3] Austin M 2007 Species distribution models and ecological theory: a critical assessment and some possible new approaches *Ecol. Modelling* **200** 1–19
[4] Barbet-Massin M, Jiguet F, Albert C H and Thuiller W 2012 Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* **3** 327–38
[5] Bartholdi J J, Jarumaneeroj P and Ramudhin A 2016 A new connectivity index for container ports *Marit. Econ. Logist.* **18** 231–49
[6] Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher S P, Peterson A T, Soberón J and Villalobos F 2011 The crucial role of the accessible area in ecological niche modeling and species distribution modeling *Ecol. Modelling* **222** 1810–9
[7] Bazzichetto M, Malavasi M, Bartak V, Acosta A T R, Rocchini D and Carranza M L 2018 Plant invasion risk: a quest for invasive species distribution modelling in managing protected areas *Ecol. Indic.* **95** 311–9
[8] Beck J, Böller M, Erhardt A and Schwanghart W 2014 Spatial bias in the GBIF database and its effect on modeling species' geographic distributions *Ecol. Inform.* **19** 10–15
[9] Benedek J and Ivan K 2018 Remote sensing based assessment of variation of spatial disparities *Geogr. Tech.* **13** 1–9
[10] Bhandari L and Roychowdhury K 2011 Night lights and economic activity in India: a study using DMSP-OLS night time images *Proc. Asia-Pacific Adv. Netw.* **32** 218
[11] Botella C, Joly A, Monestiez P, Bonnet P and Munoz F 2020 Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection *PLoS One* **15** e0232078
[12] Buchhorn M, Smets B, Bertels L, De Roo B, Lesiv M, Tsendbazar N E, Herold M and Fritz S 2020 Copernicus global land service: land cover 100 m: collection 3: epoch 2019: globe (available at: https://zenodo.org/record/3939050#.YHmC3-gzb-g)
[13] Cerqueira V, Torgo L and Mozetič I 2020 Evaluating time series forecasting models: an empirical study on performance estimation methods *Mach. Learn.* **109** 1997–2028
[14] Charles H and Dukes J S 2008 Impacts of invasive species on ecosystem services *Biological Invasions* (Berlin: Springer) pp 217–37
[15] Cook D C 2008 Benefit cost analysis of an import access request *Food Policy* **33** 277–85
[16] Cook D C and Fraser R W 2008 Trade and invasive species risk mitigation: reconciling WTO compliance with maximising the gains from trade *Food Policy* **33** 176–84
[17] Diagne C, Leroy B, Vaissière A C, Gozlan R E, Roiz D, Jarić I, Salles J M, Bradshaw C J A and Courchamp F 2021 High and rising economic costs of biological invasions worldwide *Nature* **592** 571–6
[18] Elith J and Leathwick J R 2009 Species distribution models: ecological explanation and prediction across space and time *Annu. Rev. Ecol. Evol. Syst.* **40** 677–97
[19] El-Gabbas A and Dormann C F 2018 Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent *Ecography* **41** 1161–72
[20] El-Gabbas A and Dormann C F 2018 Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling *Ecol. Evol.* **8** 2196–206
[21] European Commission 2008 Towards an EU strategy on invasive species *Technical Report*
[22] European Commission Joint Research Centre 2015 GHS-POP R2015A - GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015) - OBSOLETE RELEASE (European Commission, Joint Research Centre (JRC) [Dataset] PID) (available at: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a)
[23] FAO 2017 ISPM 5 glossary of phytosanitary terms *Technical Report* (available at: www.fao.org/fileadmin/user_upload/faoterm/PDF/ISPM_05_2016_En_2017-05-25_PostCPM12_InkAm.pdf)
[24] Fernández D and Nakamura M 2015 Estimation of spatial sampling effort based on presence-only data and accessibility *Ecol. Modelling* **299** 147–55
[25] Forbes D J 2013 Multi-scale analysis of the relationship between economic statistics and DMSP-OLS night light images *GISci. Remote Sens.* **50** 483–99

[26] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Stat. Softw.* **33** 1
[27] Gallardo B 2014 Europe's top 10 invasive species: relative importance of climatic, habitat and socio-economic factors *Ethol. Ecol. Evol.* **26** 130–51
[28] Gallardo B and Aldridge D C 2013 The 'dirty dozen': socio-economic factors amplify the invasion potential of 12 high-risk aquatic invasive species in Great Britain and Ireland *J. Appl. Ecol.* **50** 757–66
[29] Greenwell B M, Boehmke B C and McCarthy A J 2018 A simple and effective model-based variable importance measure (arXiv:1805.04755)
[30] Grinnell J 1917 Field tests of theories concerning distributional control *Am. Nat.* **51** 115–28
[31] Hampf A C, Nendel C, Strey S and Strey R 2021 Biotic yield losses in the Southern Amazon, Brazil: making use of smartphone-assisted plant disease diagnosis data *Front. Plant Sci.* **12** 548
[32] Heberling J M, Miller J T, Noesgaard D, Weingart S B and Schigel D 2021 Data integration enables global biodiversity synthesis *Proc. Natl Acad. Sci.* **118** e2018093118
[33] Hengl T 2018 Clay content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (available at: https://zenodo.org/record/2525663#.YHmBeOgzb-g)
[34] Hengl T 2018 Global maps of potential natural vegetation at 1 km resolution (available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QQHCIK)
[35] Hengl T 2018 Nighttime lights PC1-4 based on the version 4 DMSP-OLS nighttime lights time series 1997–2014 (available at: https://zenodo.org/record/1458947#.YHmGzegzb-g)
[36] Hengl T 2018 Sand content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (available at: https://zenodo.org/record/2525662#.YHmAT-gzb-g)
[37] Hengl T 2018 Soil pH in $H_2O$ at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (available at: https://zenodo.org/record/2525664#.YHmAWegzb-g)
[38] Hengl T and Gupta S 2019 Soil water content (volumetric %) for 33 kPa and 1500 kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (available at: https://zenodo.org/record/2784001#.YHl_UOgzb-g)
[39] Hengl T and Wheeler I 2018 Soil organic carbon content in x 5 g/kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (available at: https://zenodo.org/record/2525553#.YHl_1egzb-g)
[40] Hengl T and Wheeler I 2018 Soil organic carbon content in x 5 g/kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (version v0.2) [data set] (available at: https://zenodo.org/record/2525553#.YHmBVOgzb-g)
[41] Hengl T, Walsh M G, Sanderman J, Wheeler I, Harrison S P and Prentice I C 2018 Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential *PeerJ* **6** e5457
[42] Hudgins E J, Liebhold A M and Leung B 2017 Predicting the spread of all invasive forest pests in the United States *Ecol. Lett.* **20** 426–35
[43] Hulme P E 2009 Trade, transport and trouble: managing invasive species pathways in an era of globalization *J. Appl. Ecol.* **46** 10–18
[44] Hulme P E 2021 Unwelcome exchange: international trade as a direct and indirect driver of biological invasions worldwide *One Earth* **4** 666–79
[45] Hulme P E, Pyšek P, Nentwig W and Vilà M 2009 Will threat of biological invasions unite the European Union? *Science* **324** 40–41
[46] Jarnevich C S, Stohlgren T J, Kumar S, Morisette J T and Holcombe T R 2015 Caveats for correlative species distribution modeling *Ecol. Inform.* **29** 6–15
[47] Karger D N, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza R W, Zimmermann N E, Linder H P and Kessler M 2017 Climatologies at high resolution for the earth's land surface areas *Sci. Data* **4** 170122
[48] Kearney M and Porter W 2009 Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges *Ecol. Lett.* **12** 334–50
[49] Keller R P, Geist J, Jeschke J M and Kühn I 2011 Invasive species in Europe: ecology, status and policy *Environ. Sci. Eur.* **23** 23
[50] Kennedy C M, Oakleaf J R, Theobald D M, Baruch-Mordo S and Kiesecker J 2019 Managing the middle: a shift in conservation priorities based on the global human modification gradient *Glob. Change Biol.* **25** 811–26
[51] Kettunen M, Genovesi P, Gollasch S, Pagad S, Starfinger U, Ten Brink P and Shine C 2008 Technical support to EU strategy on invasive species (IAS)—assessment of the impacts of IAS in Europe and the EU (final module report for the European Commission) *Technical Report* (Institute for European Environmental Policy)
[52] Kosmala M, Wiggins A, Swanson A and Simmons B 2016 Assessing data quality in citizen science *Front. Ecol. Environ.* **14** 551–60
[53] Kulhanek S A, Leung B and Ricciardi A 2011 Using ecological niche models to predict the abundance and impact of invasive species: application to the common carp *Ecol. Appl.* **21** 203–13
[54] Leroy B, Delsol R, Hugueny B, Meynard C N, Barhoumi C, Barbet-Massin M and Bellard C 2018 Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance *J. Biogeogr.* **45** 1994–2002
[55] Levine J M, Vila M, Antonio C M D, Dukes J S, Grigulis K and Lavorel S 2003 Mechanisms underlying the impacts of exotic plant invasions *Proc. R. Soc.* B **270** 775–81
[56] Lobo J M, Jiménez-Valverde A and Hortal J 2010 The uncertain nature of absences and their importance in species distribution modelling *Ecography* **33** 103–14
[57] Luna S *et al* 2018 Developing mobile applications for environmental and biodiversity citizen science: considerations and recommendations *Multimedia Tools and Applications for Environmental & Biodiversity Informatics* (Berlin: Springer) pp 9–30
[58] Mackenzie D I and Royle J A 2005 Designing occupancy studies: general advice and allocating survey effort *J. Appl. Ecol.* **42** 1105–14
[59] Maron J L and Vilà M 2001 When do herbivores affect plant invasion? Evidence for the natural enemies and biotic resistance hypotheses *Oikos* **95** 361–73
[60] McPherson J M, Jetz W and Rogers D J 2004 The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.* **41** 811–23
[61] Meijer J R, Huijbregts M A J, Schotten K C G J and Schipper A M 2018 Global patterns of current and future road infrastructure *Environ. Res. Lett.* **13** 064006
[62] Merow C *et al* 2014 What do we gain from simplicity versus complexity in species distribution models? *Ecography* **37** 1267–81
[63] Newbold T *et al* 2016 Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment *Science* **353** 288–91
[64] O'Grady M J, Muldoon C, Carr D, Wan J, Kroon B and O'Hare G M 2016 Intelligent sensing for citizen science *Mob. Netw. Appl.* **21** 375–85
[65] Perrings C, Burgiel S, Lonsdale M, Mooney H and Williamson M 2010 Int. cooperation in the solution to trade-related invasive species risks *Ann. New York Acad. Sci.* **1195** 198–212
[66] Peterson A T, Soberón J, Pearson R G, Anderson R P, Martínez-Meyer E, Nakamura M and Araújo M B 2011 *Ecological Niches and Geographic Distributions* (Princeton, NJ: Princeton University Press)
[67] Phillips S J 2008 Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al (2007) *Ecography* **31** 272–8

[68] Phillips S J, Dudík M, Elith J, Graham C H, Lehmann A, Leathwick J and Ferrier S 2009 Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data *Ecol. Appl.* **19** 181–97
[69] Pimentel D, Zuniga R and Morrison D 2005 Update on the environmental and economic costs associated with alien-invasive species in the United States *Ecol. Econ.* **52** 273–88
[70] Renner I W, Elith J, Baddeley A, Fithian W, Hastie T, Phillips S J, Popovic G and Warton D I 2015 Point process models for presence-only analysis *Methods Ecol. Evol.* **6** 366–79
[71] Roberts D R *et al* 2017 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure *Ecography* **40** 913–29
[72] Robinet C, Roques A, Pan H, Fang G, Ye J, Zhang Y and Sun J 2009 Role of human-mediated dispersal in the spread of the pinewood nematode in China *PLoS One* **4** e4646
[73] Schneider K, Van der Werf W, Cendoya M, Mourits M, Navas-Cortés J A, Vicent A and Lansink A O 2020 Impact of *Xylella fastidiosa* subspecies *pauca* in European olives *Proc. Natl Acad. Sci.* **117** 9250–9
[74] Tabak M A, Piaggio A J, Miller R S, Sweitzer R A and Ernest H B 2017 Anthropogenic factors predict movement of an invasive species *Ecosphere* **8** e01844
[75] VanDerWal J, Shoo L P, Graham C and Williams S E 2009 Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecol. Modelling* **220** 589–94
[76] Venette R C *et al* 2010 Pest risk maps for invasive alien species: a roadmap for improvement *BioScience* **60** 349–62
[77] Venter O *et al* 2016 Global terrestrial human footprint maps for 1993 and 2009 *Sci. Data* **3** 160067
[78] Vollering J, Halvorsen R, Auestad I and Rydgren K 2019 Bunching up the background betters bias in species distribution models *Ecography* **42** 1717–27
[79] Warton D I, Renner I W and Ramp D 2013 Model-based control of observer bias for the analysis of presence-only data in ecology *PLoS One* **8** e79168
[80] Weiss D J *et al* 2018 A global map of travel time to cities to assess inequalities in accessibility in 2015 *Nature* **553** 333–6
[81] World Resource Institute 2015 Aqueduct global maps 2.1 data (available at: www.wri.org/resources/data-sets/aqueduct-global-maps-21-data)
[82] World Resources Institute 2016 *Erosion. Global Forest Watch Water. Accessed through Resource Watch, 01.02.2021* (available at: www.resourcewatch.org)
[83] Yamazaki D, Ikeshima D, Tawatari R, Yamaguchi T, O'Loughlin F, Neal J C, Sampson C C, Kanae S and Bates P D 2017 A high-accuracy map of global terrain elevations *Geophys. Res. Lett.* **44** 5844–53
[84] Zizka A *et al* 2019 CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases *Methods Ecol. Evol.* **10** 744–51
[85] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20