insects

MDPI

Article

Multi-Scenario Species Distribution Modeling

Senait D. Senay [1,2,*] and Susan P. Worner [3]

[1] GEMS™—A CFANS & MSI initiative, University of Minnesota, 305 Cargill Building, 1500 Gortner Avenue, Saint Paul, MN 55108, USA

[2] Department of Plant Pathology, University of Minnesota, 495 Borlaug Hall, 1991 Upper Buford Circle, Saint Paul, MN 55108, USA

[3] Bio-Protection Research Centre, Lincoln University, 7674 Lincoln, New Zealand; sue.worner@lincoln.ac.nz

* Correspondence: ssenay@umn.edu

Received: 8 January 2019; Accepted: 21 February 2019; Published: 1 March 2019

Abstract: Correlative species distribution models (SDMs) are increasingly being used to predict suitable insect habitats. There is also much criticism of prediction discrepancies among different SDMs for the same species and the lack of effective communication about SDM prediction uncertainty. In this paper, we undertook a factorial study to investigate the effects of various modeling components (species-training-datasets, predictor variables, dimension-reduction methods, and model types) on the accuracy of SDM predictions, with the aim of identifying sources of discrepancy and uncertainty. We found that model type was the major factor causing variation in species-distribution predictions among the various modeling components tested. We also found that different combinations of modeling components could significantly increase or decrease the performance of a model. This result indicated the importance of keeping modeling components constant for comparing a given SDM result. With all modeling components, constant, machine-learning models seem to outperform other model types. We also found that, on average, the Hierarchical Non-Linear Principal Components Analysis dimension-reduction method improved model performance more than other methods tested. We also found that the widely used confusion-matrix-based model-performance indices such as the area under the receiving operating characteristic curve (AUC), sensitivity, and Kappa do not necessarily help select the best model from a set of models if variation in performance is not large. To conclude, model result discrepancies do not necessarily suggest lack of robustness in correlative modeling as they can also occur due to inappropriate selection of modeling components. In addition, more research on model performance evaluation is required for developing robust and sensitive model evaluation methods. Undertaking multi-scenario species-distribution modeling, where possible, is likely to mitigate errors arising from inappropriate modeling components selection, and provide end users with better information on the resulting model prediction uncertainty.

Keywords: invasive insect species; model uncertainty; multi-model framework; non-linear principal component analysis; principal component analysis; random forest; species distribution models

1. Introduction

Various species distribution models have been used to predict suitable insect habitats. Many studies are based on correlative models that use species presence along with environmental data to infer suitable habitats for the species under study [1]. Currently, discrepancies between model results represent a major issue in ecological modeling, and the need for quantifying model uncertainty has been repeatedly discussed in the literature [2–12]. A few important studies have investigated sources of uncertainty in SDMs [6,7,9,10,12], and some have developed techniques for quantifying the uncertainty associated with modeling species' distribution in ecology and the wider spatial modeling context [6,13,14].

Variation in the predictive accuracy of different models is usually attributed to the inherent robustness of modeling algorithms. Simple models, for example, the bioclimatic analysis and prediction system, BIOCLIM [15] and the point-to-point similarity metric system, DOMAIN [16], have been reported to be suitable for predicting the distribution of rare species occupying a limited environmental niche, representing simple linear interactions among environmental variables. Complex models, such as Support Vector Machines (SVMs) [17] and Artificial Neural Networks (ANNs), with complex functions that consider non-linearity and a large number of variables can handle complex interactions within a multidimensional variable space [5,18]. Jiménez-Valverde, et al. [19] and Chefaoui and Lobo [20] argued that the above claim is not entirely true due to inappropriate comparison of model prediction results for species with varying relative occurrence areas in the above studies. However, there is general agreement of an inherent difference in the predictions of simple and complex models based on differences between model predictions in studies that compare models using the same occurrence data and study area [21].

The use of geo-environmental variables in addition to purely climatic variables such as temperature and precipitation, has been reported to increase prediction accuracy in SDM models [1,22–24]. However, many studies still depend on using a limited number and type of variables, for example, variables derived only from temperature and precipitation data without additional climatic or geo-environmental information, such as elevation, that might help a model better discriminate an ecological niche [25]. Often, the spatial variation of geo-environmental variables is much higher than climatic variables, and can help characterize unique habitats when used along with climatic variables [24]. General reluctance to using additional environmental variables may be associated with data inconsistency that can occur when using data compiled from different sources and multi-scale variables. As the types and number of variables increase, the likelihood of obtaining variables from multiple sources (e.g., sensors) increases. Moreover, the scales of multi-source variables are also likely to differ, thereby significantly increasing modeling effort. Another, significant complication from increasing the number of variables is the increasing complexity of interactions among large sets of variables. Climatic variables frequently used in a number of species' distribution studies are assumed to have linear relationships. However, linear relationships are not always observed between environmental variables, especially if large sets of predictors from multiple scales and data sources are used in the modeling process. Therefore, the type and number of variables used can further increase the divergence of predictions between simple and complex models because of interactions among environmental variables used [10,19]. Moreover, any dimension reduction performed on predictor datasets may also affect the predictive accuracy of SDMs. Numerous dimension-reduction methods have been used for various applications, but only a few have been successfully adopted in ecology [9,26,27]. When using multi-sourced environmental predictor variables in combination with climatic variables, it is important to consider the possibility of complex and non-linear interactions between variables before choosing a dimension reduction method. For instance, a multidimensional predictor dataset that may not be adequately represented by a simple model may be successfully modeled after application of appropriate dimension-reduction methods [28]. Therefore, in addition to differences in the nature of the species-occurrence data, choice of predictors, and the models used for prediction, data pre-processing methods, such as dimension-reduction techniques or the method used for variable selection, could affect the performance of a SDM prediction.

The purpose of this study was to investigate the effect of modeling components such as predictor datasets, dimension-reduction methods, and model types, and their interactions on SDM model performance using a factorial experimental design and selected case studies. We further investigated whether it is possible to increase model accuracy by using appropriate predictors, dimension-reduction methods, and model types that better explain the spatial distribution of presence points, and their pattern in the predictor feature space. Finally, we discuss the advantage of using a multi-scenario modeling framework to provide information about model prediction uncertainty.

2. Materials and Methods

2.1. Geographic Extent of Study Area

This study was carried out over a global scale, therefore, all environmental predictors had global coverage. The different species-presence datasets, however, cover varying spatial extents. All datasets have a resolution of 10 arc minutes (0.17°, 18.5 km at the equator). The secondary geographic focus of the study was the extent covered by the North Island, South Island, and Stewart Island of New Zealand.

2.2. Predictor Data

Three predictor datasets were used (Table 1), 1) BIOCLIM19 (P1) consisting of 19 temperature- and precipitation-related variables derived from the WORLDCLIM dataset [29,30]; 2) BIOCLIM35 (P2) consisting of the BIOCLIM19 variables and additional 16 radiation and water-balance soil moisture index-related variables accessed from the CLIMOND dataset [31]; and, 3) the BIOCLIM35+T4 (P3). In P3 there was a set of four topographic variables (elevation, slope, aspect, and hill-shade) derived from a digital elevation model (DEM), downloaded from the WORLDCLIM data portal [29,32], which were added to the P2 dataset. The variables, slope, aspect, and hill-shade were calculated from the elevation data using ESRI's ArcInfo® spatial analyst software. A 3×3 pixel focal area was used to process all three DEM-derived topographical variables. Detailed information on the development of variables in the P1 and P2 datasets is given in Hijmans, et al. [32] and Kriticos, et al. [31], respectively.

2.3. Dimension Reduction

Three dimension-reduction methods were used: first, a variable selection with the random forest algorithm (RF; DR1). The RF algorithm can handle large numbers of variables and it is widely used in species distribution modeling. The random forest classifier results in low-bias selection by averaging over a large ensemble of high-variance but low-correlation trees [33]. The Akaike information criterion (AIC) was used to rank variable importance; second, principal component analysis (PCA; DR2) was used. PCA is a mathematical method that transforms a set of raw variables into linearly uncorrelated variables by mapping the newly transformed data on artificial orthogonal axes [34]. The PCA itself, or slightly modified versions of it, have been used in ecological modeling either as a dimension-reduction method or as the main species distribution model [35,36]. The third dimension reduction method was non-linear principal component analysis (NLPCA; DR3) in the form of a hierarchical NLPCA (h-NLPCA), a neural network model developed by Scholz and Vigario [37]. The method is reported to be the true non-linear extension of the linear PCA [37]. The h-NLPCA achieves a hierarchical order of principal components similar to the linear PCA (Appendix A) and is both scalable and stable similar to the linear PCA method [38]. Most h-NLPCA parameters were internally computed as they are adjusted throughout the iterative learning. For this study, network weights were initialized at random. The weight decay was set at 0.001 with the maximum iteration conditionally set by either five times the number of observations or 3000, whichever was minimum.

Table 1. Variables included in the three-predictor datasets used in this study.

Variable	Variable Name	Dataset
01	Annual mean temperature (°C)	P1, P2, P3
02	Mean diurnal temperature range (mean(period max-min)) (°C)	P1, P2, P3
03	Isothermality (Bio02 ÷ Bio07)	P1, P2, P3
04	Temperature seasonality (C of V)	P1, P2, P3
05	Max temperature of warmest week (°C)	P1, P2, P3
06	Min temperature of coldest week (°C)	P1, P2, P3
07	Temperature annual range (Bio05-Bio06) (°C)	P1, P2, P3
08	Mean temperature of wettest quarter (°C)	P1, P2, P3
09	Mean temperature of driest quarter (°C)	P1, P2, P3
10	Mean temperature of warmest quarter (°C)	P1, P2, P3
11	Mean temperature of coldest quarter (°C)	P1, P2, P3
12	Annual precipitation (mm)	P1, P2, P3
13	Precipitation of wettest week (mm)	P1, P2, P3
14	Precipitation of driest week (mm)	P1, P2, P3
15	Precipitation seasonality (C of V)	P1, P2, P3
16	Precipitation of wettest quarter (mm)	P1, P2, P3
17	Precipitation of driest quarter (mm)	P1, P2, P3
18	Precipitation of warmest quarter (mm)	P1, P2, P3
19	Precipitation of coldest quarter (mm)	P1, P2, P3
20	Annual mean radiation (W m^{-2})	P2, P3
21	Highest weekly radiation (W m^{-2})	P2, P3
22	Lowest weekly radiation (W m^{-2})	P2, P3
23	Radiation seasonality (C of V)	P2, P3
24	Radiation of wettest quarter (W m^{-2})	P2, P3
25	Radiation of driest quarter (W m^{-2})	P2, P3
26	Radiation of warmest quarter (W m^{-2})	P2, P3
27	Radiation of coldest quarter (W m^{-2})	P2, P3
28	Annual mean moisture index	P2, P3
29	Highest weekly moisture index	P2, P3
30	Lowest weekly moisture index	P2, P3
31	Moisture index seasonality (C of V)	P2, P3
32	Mean moisture index of wettest quarter	P2, P3
33	Mean moisture index of driest quarter	P2, P3
34	Mean moisture index of warmest quarter	P2, P3
35	Mean moisture index of coldest quarter	P2, P3
36	Elevation (m)	P3
37	Slope (deg)	P3
38	Aspect (deg)	P3
39	Hillshade	P3

2.4. Species Data

The worldwide distribution of five insect species: (i) *Aedes albopictus* (Skuse, 1894), (ii) *Anoplopis gracilipes* (Smith, 1857), (iii) *Diabrotica virgifera virgifera* (LeConte, 1868), (iv) *Thaumetopoea pityocampa* (Denis & Schiffermuller, 1775), and (v) *Vespula vulgaris* (Linnaeus, 1758) (established in New Zealand), was compiled from three sources. The sources included: (a) the Global Biodiversity Information Facility (GBIF) database, (b) previous literature, and (c) personal communication with domain experts (Figure 1). The geographical extents covered by the presence locations for these species vary widely. Variation in the relative occurrence area (ROA) among these species was important to investigate if the distribution range of a species affects the predictive accuracy of SDMs. *A. albopictus* and *V. vulgaris* have a relatively large ROA (Figure 1A,F), whereas, *D. v. virgifera* and *A. gracilipes* cover an intermediate global extent (Figure 1B,C). *T. pityocampa* (Figure 1E) has the smallest occurrence cover, hence, the smallest ROA (See Note S5 for description on the native and invaded ranges of these five insect species).

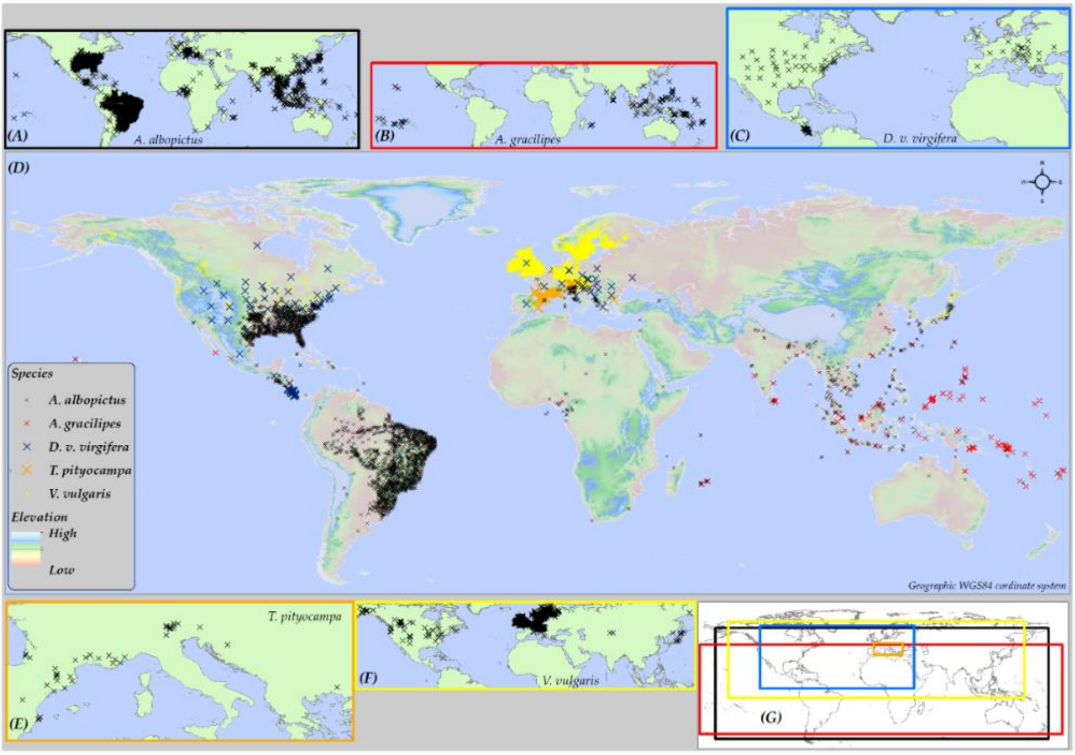

Figure 1. Maps showing the global occurrence of the five species used in this study. Inset maps show the global distribution of (**A**) *Aedes albopictus*; (**B**) *Anoplopis gracilipes*; (**C**) *D. v. virgifera*; (**E**) *Thaumetopoea pityocampa*; and (**F**) *Vespula vulgaris*. (**G**) The geographic extent of each species; and main map (**D**) shows the extent of occurrence of all five species with presence points overlaid on a global elevation model.

Pseudo-absences were generated using a 3-step pseudo-absence generation method [21]. The first step involves choosing a geographical distance to bind or constrain background data from which pseudo-absences were selected. A measure of how far the correlation structure among variables is conserved through the environmental space using presence points as a reference is used to determine the appropriate distance. The distance at which variable importance changes is used as an indicator of correlation structure change. Variable importance ranking is done by carrying out PCA on successive datasets extracted at incremental geographical distances from presence points [21]. At the second step, environmental profiling is used to discriminate environmentally dissimilar portions of the background dataset bound at a distance identified in step one. Lastly, in the third step, a number of pseudo-absences that balance the number of presences, are selected by k-means clustering the class that is most environmentally dissimilar from presences in step two (refer to Senay, et al. [21] for the details on how to generate pseudo absences using the 3-step pseudo-absence selection method. Iturbide, et al. [39] describe a slightly different methodology). The correlation structure among variables changes as the occurrence points-predictor data combinations change making the variable importance over distance different [21]. Therefore, three distances were calculated (Table 2) for each species modeled using the three different predictor datasets (P1, P2, and P3).

Table 2. Number of presence points * and distances used to limit background extent before pseudo-absence selection, for the three types of predictor datasets used to model the global distribution of the five species in this study.

No.	Species	Predictor	Distance (km)
1	*Aedes albopictus* (3029/2928)	BIOCLIM19	350
2	*Aedes albopictus* (3029/2928)	BIOCLIM35	300
3	*Aedes albopictus* (3029/2928)	BIOCLIM35+T4	600
4	*Anoplopis gracilipes* (385/101)	BIOCLIM19	550
5	*Anoplopis gracilipes* (385/101)	BIOCLIM35	500
6	*Anoplopis gracilipes* (385/101)	BIOCLIM35+T4	400
7	*Diabrotica v. virgifera* (449/84)	BIOCLIM19	2000
8	*Diabrotica v. virgifera* (449/84)	BIOCLIM35	800
9	*Diabrotica v. virgifera* (449/84)	BIOCLIM35+T4	800
10	*Thaumetopoea pityocampa* (67/33)	BIOCLIM19	300
11	*Thaumetopoea pityocampa* (67/33)	BIOCLIM35	1300
12	*Thaumetopoea pityocampa* (67/33)	BIOCLIM35+T4	800
13	*Vespula vulgaris* (10,048/920)	BIOCLIM19	550
14	*Vespula vulgaris* (10,048/920)	BIOCLIM35	300
15	*Vespula vulgaris* (10,048/920)	BIOCLIM35+T4	700

* Numbers next to each species name show available presence points followed by spatially unique points with respect to the environmental predictor dataset resolution. Another two sets of the datasets listed above were generated according to the listed background binding distances for predictor data transformed using PCA and NLPCA making 45 training/test datasets in total. BIOCLIM19 contains 19 precipitation- and temperature-based variables, BIOCLIM35 contains variables in BIOCLIM19 plus 26 radiation- and soil moisture-derived variables, BIOCLIM35+T4 contains variables in BIOCLIM35 plus 4 variables derived from topographic data.

Pseudo-absence selection was carried out after the background data was projected onto the Mercator projected coordinate system and interpolated to an equal area grid. The equal area grid projected data was chosen so that any poleward bias that might occur by selecting pseudo-absences within a buffer defined in the unequal grid geographic coordinate system is avoided. The selected pseudo-absence points were re-projected onto a geographic coordinate system for the remainder of the modeling process (Figure 2).

2.5. Model Type

An exhaustive comparison of all available SDMs would be impractical. In this study, we selected four model types to provide predictions for the global distribution of each species based on various P and DR combinations. The first model is quantitative discriminant analysis—QDA (MT1): discriminant analyses in general, and QDA in particular, are classical multivariate models used in species distribution modeling [10,40,41]. While QDA allows easy assessment of variable contributions and assessment of the species distribution prediction, it cannot handle datasets where the number of observations is smaller than the number of variables. The qda function from R [42] and MASS [43] libraries was used to run QDA in the multi-model framework. The second was logistic regression-LOGR (MT2). LOGR is one of the most frequently used SDMs for species distribution studies [6,44]. The glm function in the Stats [42] package in R was used to run LOGR in the multi-model framework. The third model was classification and regression trees—CART (MT3). CART is a classification and regression decision tree that is also frequently used in species distribution models [40,45]. It has been suggested that decision trees incorporate the complexity needed to explain interactions between multidimensional variable data, without a complicated rule that can be easily explained to end-users [40]. The rpart package for R was used to run CART in the multi-model framework. The fourth was a support vector machine—SVM (MT4). Support vector machines are models based on machine learning theory, specifically, artificial neural networks [17]. SVM has been shown to be an excellent classifier in a number of disciplines, for example, in astronomy, medicine, physics, and pattern recognition [46]. SVM has recently been used in ecological modeling along with other machine learning methods such as boosted regression trees (BRT) and artificial neural networks

(ANNs) [40,44]. The SVM model has the option of fitting data either linearly or non-linearly. The specific functions used were linear, radial basis, and polynomial. The Kernlab [47] package for R was used to run SVM in the multi-model framework. A separate optimization of the model was carried out to identify the best parameters for the different datasets. The multi-model framework developed by Worner, et al. [41] was used to run the four models in a standardized set-up. Model parameterization and data exporting was also done using this framework.

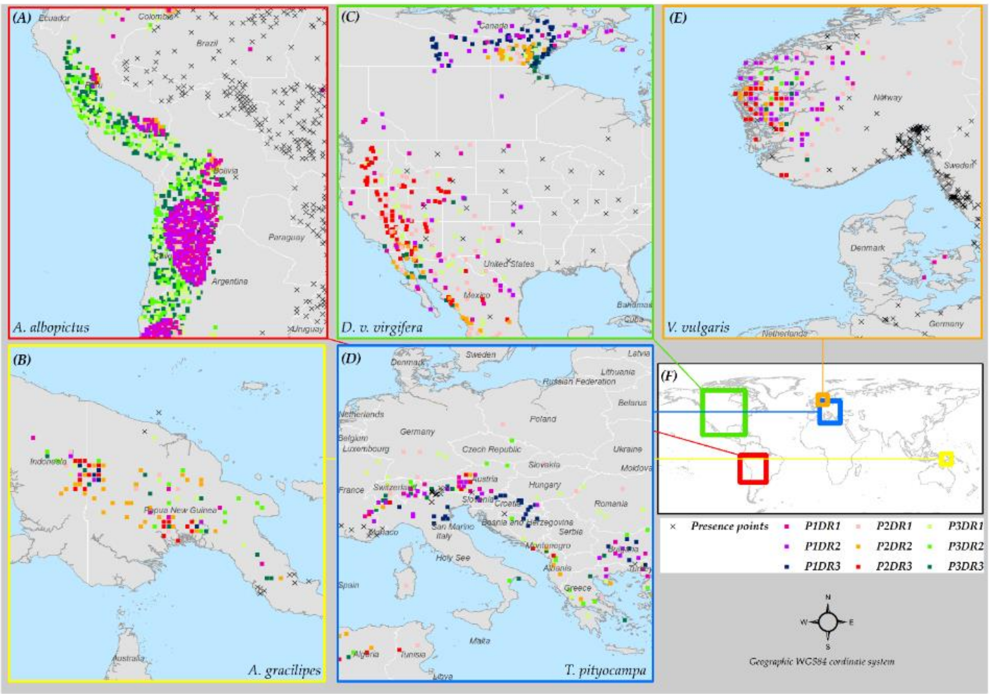

Figure 2. Subsets of the global study area with different sets of pseudo-absence points shown by the species, predictor data, and dimension reduction method. The nine sets of pseudo-absences generated based on the different combinations of the three predictor datasets and three dimension reduction methods are shown for *A. albopictus* (**A**), *A. gracilipes* (**B**), *D. v. virgifera* (**C**), *T. pityocampa* (**D**), and *V. vulgaris* (**E**). The extents of the sub-set maps (**A**–**E**) are shown on the global map (**F**).

2.6. Research Design and Model Conceptualization

Presence data for five species (PD) were used, and sets of pseudo-absence (PA) data were developed for three predictor datasets (P) and three dimension reduction methods (DR) comprising, 5 PD × 3 P × 3 DR = 45 PA datasets (Figure 3). Finally, 45 training datasets (TD) were prepared by combining PD and PA datasets for each PD-P-DR combination. TD datasets for each species were used to train and evaluate four different models (MT). Within each TD dataset, 80% of the data was used for model training and 20% for model evaluation. Each run was replicated 20 times. Each MT was used to predict the global distribution of each species based on the various P and DR combinations comprising, 3 P × 3 DR × 4 MT = 36 different predictions for each species (Figure 3). Model-performance metrics were calculated and used to evaluate the performance of each model-prediction.

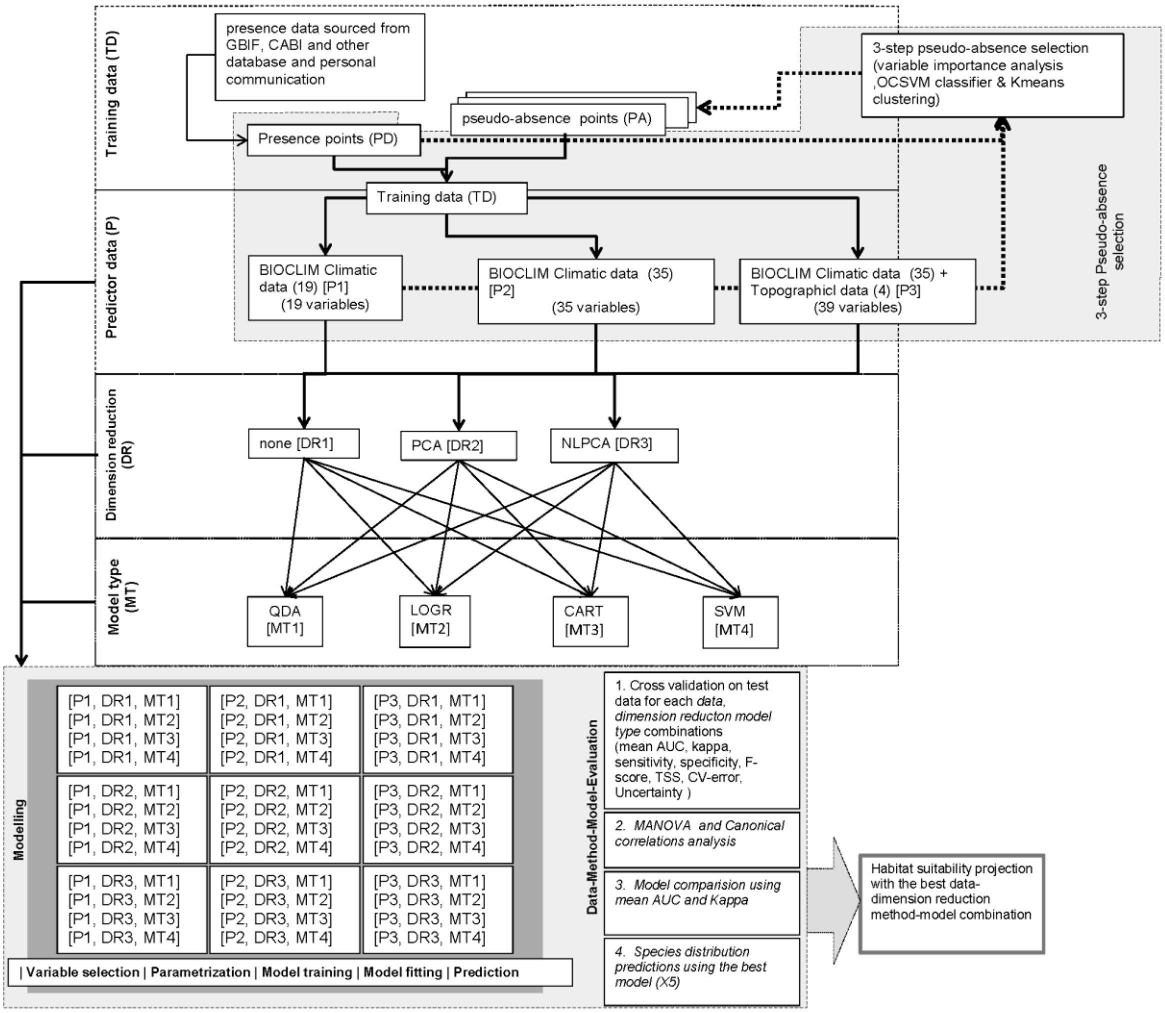
Training data (TD)
presence data sourced from GBIF, CABI and other database and personal communication
Presence points (PD)
pseudo-absence points (PA)
3-step pseudo-absence selection (variable importance analysis ,OCSVM classifier & Kmeans clustering)
3-step Pseudo-absence selection
Training data (TD)
Predictor data (P)
BIOCLIM Climatic data (19) [P1] (19 variables)
BIOCLIM Climatic data (35) [P2] (35 variables)
BIOCLIM Climatic data (35) + Topographicl data (4) [P3] (39 variables)
Dimension reduction (DR)
none [DR1]
PCA [DR2]
NLPCA [DR3]
Model type (MT)
QDA [MT1]
LOGR [MT2]
CART [MT3]
SVM [MT4]
Modelling
[P1, DR1, MT1]
[P1, DR1, MT2]
[P1, DR1, MT3]
[P1, DR1, MT4]
[P2, DR1, MT1]
[P2, DR1, MT2]
[P2, DR1, MT3]
[P2, DR1, MT4]
[P3, DR1, MT1]
[P3, DR1, MT2]
[P3, DR1, MT3]
[P3, DR1, MT4]
[P1, DR2, MT1]
[P1, DR2, MT2]
[P1, DR2, MT3]
[P1, DR2, MT4]
[P2, DR2, MT1]
[P2, DR2, MT2]
[P2, DR2, MT3]
[P2, DR2, MT4]
[P3, DR2, MT1]
[P3, DR2, MT2]
[P3, DR2, MT3]
[P3, DR2, MT4]
[P1, DR3, MT1]
[P1, DR3, MT2]
[P1, DR3, MT3]
[P1, DR3, MT4]
[P2, DR3, MT1]
[P2, DR3, MT2]
[P2, DR3, MT3]
[P2, DR3, MT4]
[P3, DR3, MT1]
[P3, DR3, MT2]
[P3, DR3, MT3]
[P3, DR3, MT4]
| Variable selection | Parametrization | Model training | Model fitting | Prediction
Data-Method-Model-Evaluation
1. Cross validation on test data for each data, dimension reducton model type combinations (mean AUC, kappa, sensitivity, specificity, F-score, TSS, CV-error, Uncertainty)
2. MANOVA and Canonical correlations analysis
3. Model comparision using mean AUC and Kappa
4. Species distribution predictions using the best model (X5)
Habitat suitability projection with the best data-dimension reduction method-model combination

Figure 3. Conceptual model showing factorial research design. The study was carried out using a 3 × 5 × 3 × 4 factorial design. The design incorporated three types of predictor datasets, occurrence data for five species, three types of collinearity reduction methods, and four types of models that utilize different modeling techniques.

2.7. Model Choice Evaluation

A multiple factor multivariate analysis of variance (MANOVA) was carried out to investigate the effect of TD, P, DR, and MT on five metrics commonly used to measure model performance. Kappa [48] developed by Cohen [49], area under the receiver operating characteristic curve (AUC) [50], sensitivity [48,51], specificity [48,51], and cross-validation error [52] given by the root mean square error (RMSE) of the variations in cross-validation iterations. MANOVA was used to obtain a statistically informed decision regarding which performance measure to use for evaluating factorial design results. Mahalanobis distances [53] derived between each performance score and the group (P, DR and MT) centroids were compared using the Chi-square (x^2) distribution and plotted on a q–q plot [54], to determine multivariate normality of the model's performance scores. The majority of the data conformed to the expected χ^2 value with a few outliers. According to Box's M test [55], the equality of variance assumption was fulfilled for P and MT, but not for TD and DR. We proceeded with the parametric MANOVA test considering the fact that none of the largest standard deviations within a group (factors) were more than four times larger than the smallest standard deviations within the same group, which suggested that MANOVA will still be robust [56]. As a cautionary measure, a conservative α value for the variables that failed the Box's M test was used while carrying out the follow-up post-hoc group comparisons (i.e., α = 0.025 instead of the usual α = 0.05). Canonical correlation analysis was used to determine model performance measures that best described the effects of the modeling components. The standardized coefficients of the canonical correlation analysis were used to select the best model performance measures. The effects of predictor data type, dimension reduction, and model type on individual model performance indices were analyzed using single factor analysis of variance (ANOVA) with Tukey's honestly significant difference (HSD) post-hoc test. The multivariate statistical analysis was carried out in R [42] statistical software version 3.0.2, and with packages agricolae [57], candisc [58], CCA [59], ggplot2 [60], heplots [61], hier.part [62], and multcomp [63]. The model with the maximum score for the chosen model performance measure based on the MANOVA analysis was ranked the best model, and the model with the lowest score was considered the worst model. For *V. vulgaris* predictions, external validation was undertaken as additional occurrence data was obtained after modeling was completed. Hierarchical partitioning [64,65] was carried out to quantify the independent contribution of modeling factors, TD, P, DR, and MT to mean Kappa and cross validation scores.

Individual variables were ranked based on the frequency of their inclusion in the tested models. The method described by Dormann, et al. [9] was adapted for this purpose. The frequency of variable inclusion was calculated based on the number of times a variable was used by a model regardless of whether it was by a straight forward variable selection (RF) or by dimension reduction (PCA, and NLPCA).

Finally, the variability between the predictions across the 36 scenarios for each species were analyzed. Mean and standard deviation maps were also produced so that the spatial pattern of variability among the models can be easily visualized. The probability density of standard deviation by modeling components (P, DR, and MT) were plotted to investigate if any variation from the multivariate analysis of model performance scores was also reflected in the predicted data.

3. Results

3.1. Multivariate Analysis of Variance (MANOVA)

The MANOVA results (Table 3) showed that all modeling components and their interactions had a significant effect on the linear combination of the five model performance scores (Kappa, AUC, sensitivity, specificity, and CV-error) with the exception of predictor choice (P).

Table 3. The multiple factor multivariate analysis of variance (MANOVA) results table showing the effects of the various modeling components model performance.

Modeling Components	Pillai's Trace	η^2 (%)	F	Df	*P* [#]
Model type	0.79	26.22	9.24	3	<0.001 ***
Dimension reduction	0.42	21.01	6.86	2	<0.001 ***
Species data	0.81	20.32	6.68	4	<0.001 ***
Predictor	0.11	5.50	1.50	2	0.138 ^{ns}
Species data x Predictor	0.68	13.51	2.58	8	<0.001 ***
Species data x Dimension reduction	0.58	11.65	2.18	8	<0.001 ***
Predictor x Dimension reduction	0.49	12.37	3.70	4	<0.001 ***
Species data x Predictor x Dim. Red.	0.95	18.98	1.93	16	<0.001 ***
Residuals		26.22		132	

[#] Signif. *P* codes: $0 < *** \leq 0.001 < ** \leq 0.01 < * \leq 0.05 \leq 0.1 < \text{ns} \leq 1$.

A follow up canonical correlation analysis was undertaken, and the first canonical variable accounted for 52.4% of model variance. The corresponding canonical correlation coefficient for the first variable was 0.9034 (Wilks $\lambda = 0.015$, $F = 3.53$, $DF1 = 240$, $DF2 = 638$, and $p < 0.0001$) showing strong linear correlation between the canonical factor loadings of performance measures and that of modeling components (Figure 4). Accordingly, the canonical coefficient of determination ($R^2 = 0.816$) was also high, showing that the ratio of the explained variance out of the total variance was high. The standardized coefficients of the canonical correlation analysis showed that the Kappa score contributed to most of the variance of the first canonical variable (79.9%), and cross-validation error contributed to most of the second canonical variable (62.7%). Accordingly, model Kappa scores and cross validation error were selected for all subsequent comparisons and analyses (complete results of the canonical analyses are given in Table S1.1 and Table S1.2).

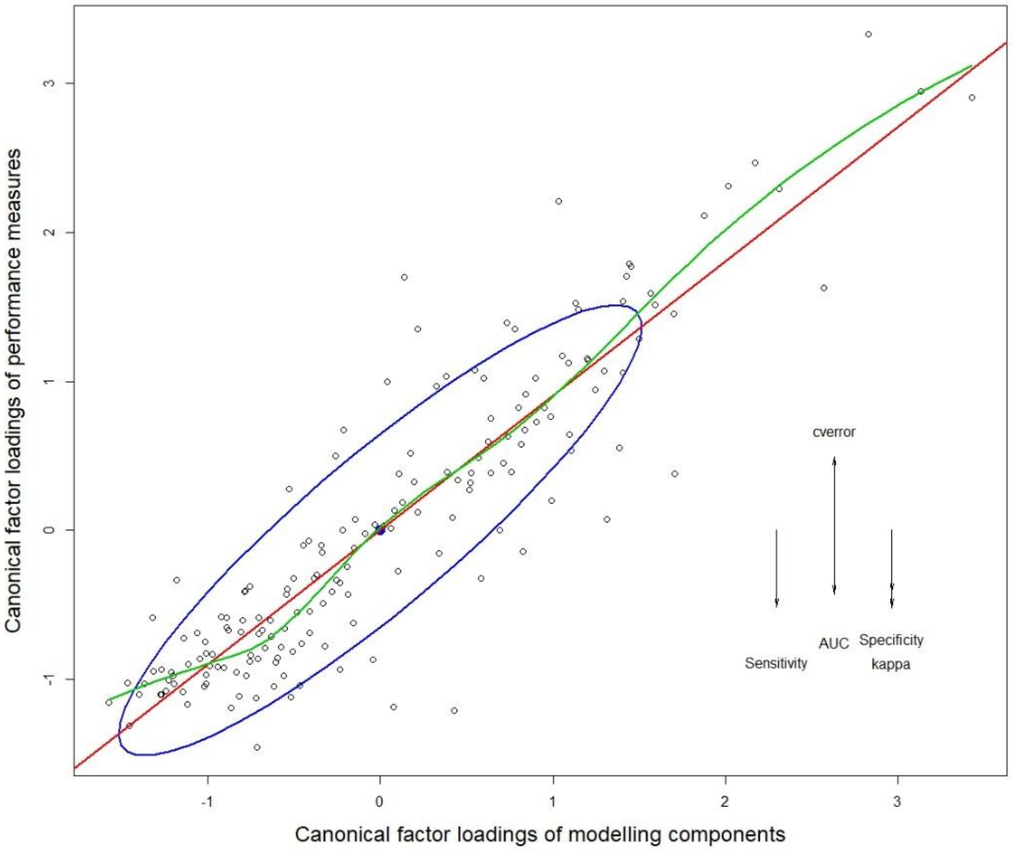

Figure 4. Structure correlations (canonical factor loadings) for the first canonical dimension. Arrows show the vector direction of variables that correspond to the canonical component on the y-axis. The corresponding variables for the x-axis (combinations of modeling components) were not labeled to avoid overcrowding the graph. The red line indicates the linear regression line. The blue ellipse (data ellipse) shows 68% of the data points (approx. one standard deviation and their centroid (filled black dot) in relation to the linear regression line. The green line shows the locally weighted scatterplot smoothing (LOWESS) fit.

3.2. *Quantifying the Variance Contribution of Modeling Factors*

Individual follow-up ANOVA's were performed for Kappa and cross-validation error scores, and the results largely agreed with the MANOVA analysis. The statistics for Kappa scores were as follows. All main effects were statistically significant (ANOVA test, SS > 0.24, $\eta^2 > 0.12$, and p 0.000–0.002), with the exception of predictor choice (SS = 0.007, $\eta^2 = 0.003$, and $p = 0.764$). Statistics for the interactions were also significant and comparable with main effects (ANOVA test, SS 0.17–0.52, η^2 0.03–0.05, and p 0.004–0.01).

The hierarchical partitioning analysis identified TD as a source of the largest variation both in Kappa scores and in model cross-validation errors (54.8% and 47.5%, respectively), followed by MT, which accounted for 38.1% and 43.8% of the variation in Kappa and CV error scores, respectively. DR accounted for 6.8% in Kappa score variation and 8.6% in cross validation error variation, and P scored 0.2% and 0.1% for Kappa and cross validation error variation, respectively.

3.3. Species Level Analysis of Variance

The species level model performance analysis showed that for *A. albopictus*, DR (ANOVA, SS = 0.118, df = 2, and p = 0.004), MT (SS = 0.689, df = 3, and p = <0.0001), and their interaction (SS = 0.259, df = 6, p = 0.001) had a significant effect on model Kappa scores. For *A. gracilipes*, only P had a significant effect on model performance scores (ANOVA, SS = 0.132, df = 2, and p = 0.004). However, pairwise comparisons of P and DR combinations (Tukey's test, HSD = 0.24, and α = 0.05) showed that PCA-based dimension reduction gave the lowest Kappa scores for *A. gracilipes* predictions. *D. v. virgifera* results were similar to *A. albopictus,* except that the interaction between DR and MT was not significant. P (ANOVA, SS = 0.212, df = 2, and p = 0.026) and MT (ANOVA, SS = 0.552, df = 3, and p = 0.001) in *T. pityocampa,* had a significant effect on model performance. Finally, for *V. vulgaris*, only MT had a significant effect (ANOVA, SS = 0.128, df = 3, and p < 0.0001).

3.4. Modeling Components

3.4.1. Species Data

A. albopictus and *V. vulgaris* had the highest mean Kappa scores, suggesting that the highest prediction accuracy ranks were associated with species that had the largest presence data records (*A. albopictus*, *V. vulgaris*). Presence data prevalence was consistently associated with high prediction accuracy when compared among the five species. For example, *A. gracilipes* which had higher presence records than *D. v. virgifera* and *T. pityocampa,* had higher Kappa scores and lower CV error than both species.

3.4.2. Predictor Choice/Variable Selection

The variables mainly included across the different modeling combinations were annual precipitation (mm), precipitation of wettest quarter (mm), precipitation of driest quarter (mm), precipitation of warmest quarter (mm), mean temperature of wettest quarter (°C), precipitation of coldest quarter (mm), and annual mean temperature (°C). The complete rank is given in Table S2.

3.4.3. Dimension Reduction

DR interacted significantly with TD, where its effect on both Kappa and cross-validation error scores varied between species datasets. NLPCA generally outperformed both PCA and RF for all species except *T. pityocampa*, where the Kappa score from RF was slightly higher (see Table S4.1 for mean Kappa scores of all TD-DR-MT combinations). This is especially true for the two species *A. albopictus* and *V. vulgaris* that had a large number of presence point records covering large environmental and geographical areas. There was a difference in Kappa scores due to DR for some of the species. For example, Kappa value increased in magnitude by0.25 for the *D. v. virgifera* distribution model when using NLPCA compared with RF. RF had a higher Kappa and lower cross-validation error values compared to PCA for *T. pityocampa* and *A. albopictus*, while PCA performed better than RF for *D. v. virgifera* and *A. gracilipes*. The mean Kappa scores for RF and PCA were very similar for *V. vulgaris*. The generally poor performance of PCA reported by Dormann, et al. [9] was also observed in this study. With regard to the interaction with model types, there were no clear trends except for the LOGR model and PCA combinations, which consistently gave poorer model performance scores.

3.4.4. Model Type

Model type effect trend was consistent throughout all combinations of factors. The SVM model consistently outperformed the other three models (Table S4.2). SVM and CART models were consistently ranked with the highest Kappa score and lowest CV-error groups. LOGR had a generally low Kappa and high CV-error scores throughout the factorial combinations. There was only one instance where LOGR scored better than QDA and CART for *A. gracilipes* within the group of models using the RF variable selection method.

3.5. Model Ranking

Kappa score was used to rank model performance for the five species distribution predictions. Discriminating models that had similar Kappa scores was done by ranking them according to the cross-validation error values (Figure 5). The best and worst P-DR-MT combinations from the 36 scenarios for each species are given in Table 4. Worst prediction in this context does not imply that the reported dimension reduction or model types are not generally suited for the particular species, rather the recommendation is specific to the environmental data, presence records, and spatial extent used in this study.

There were a number of interesting results where certain combinations did worse, despite belonging to a species with high presence prevalence. For example, the worst overall kappa score belonged to a prediction based on PCA-transformed data fitted by a logistic regression model for *A. albopictus*. Despite most of the predictions for *A. albopictus* being highly ranked according to Kappa scores (five out of the top 10 ranks out of 60 combinations; Table S4.1), this particular prediction came last (60th) with a Kappa score of 0.34. When random forest variable selection was used instead of PCA, prediction for the same model (logistic regression) and species (*A. albopictus*) scored a Kappa = 0.72, which was ranked at 42 (Table S4.1).

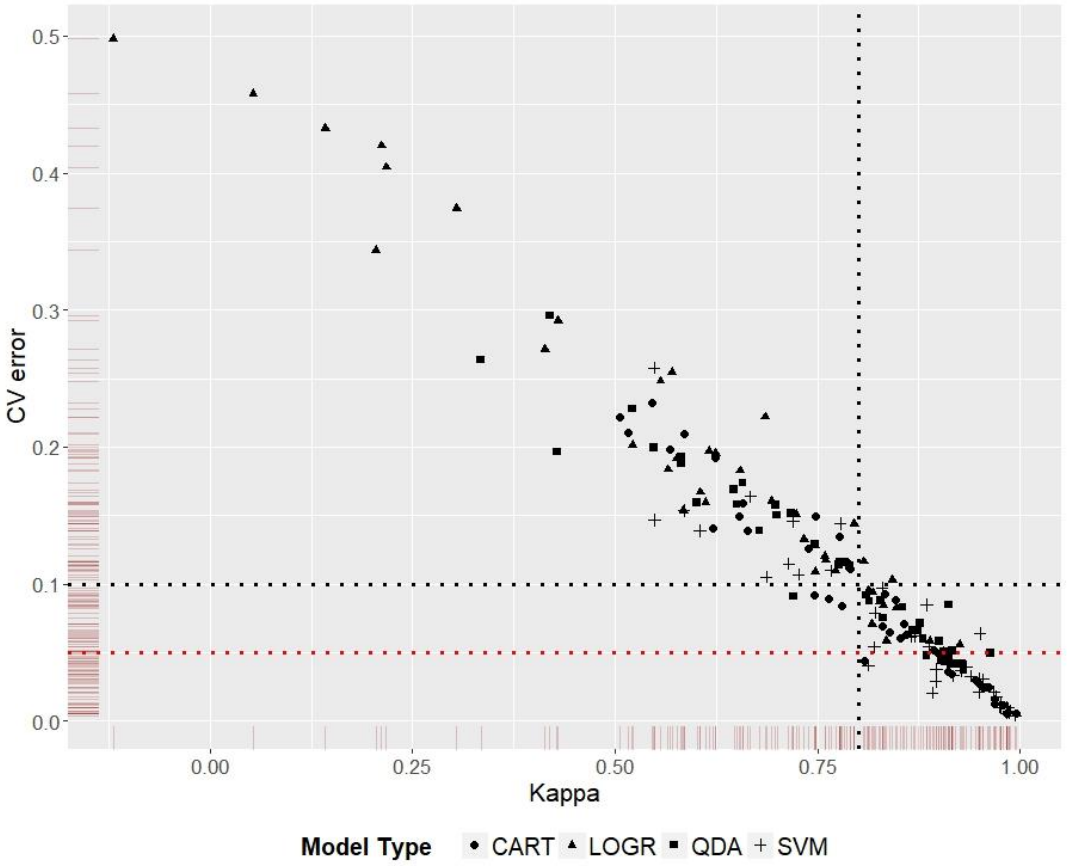

Figure 5. Model Kappa scores plotted against cross-validation error scores. Models to the right of the vertical black dotted line have a Kappa score ≥0.8; models below the horizontal black dotted line have a cross validation error ≤0.1, and models below the horizontal red dotted line have a cross validation error ≤0.05. The graph shows the advantage of using a second performance score to discriminate between models with similar scores on the first performance measure.

Table 4. Best and worst component combinations for the five species modeled in this study. For abbreviations, please refer to Figure 3.

Species	Best	Kappa	CVerror	Worst #	Kappa	CVerror
A. albopictus	P_1DR_3SVM *	0.99	0.006	P_1DR_2LOGR	0.14	0.433
A. gracilipes	P_1DR_2QDA *	0.96	0.050	P_2DR_2LOGR	0.43	0.292
D. v. virgifera	P_1DR_3SVM *	0.98	0.006	P_2DR_2LOGR	0.21	0.344
T. pityocampa	P_2DR_2SVM *	0.88	0.009	P_3DR_3LOGR	−0.12	0.498
V. vulgaris 1	P_1DR_3SVM *	0.99	0.004	P_1DR_3LOGR	0.56	0.248
V. vulgaris 2	P_1DR_3CART *	0.99	0.005			

* Combinations are the best based on their high Kappa and low CVerror, but are not significantly different from the second best combination. # All model combinations identified as "worst" for a species had a significantly lower score than the second worst models. CVerror = cross-validation error. For *V. vulgaris,* additional presence data was obtained from the Landcare Research Centre (Figure S3). External validation of the selected model for *V. vulgaris* using the New Zealand *V. vulgaris* presence data showed that 91% of the occurrence sites were correctly predicted by the selected model. Two combinations were selected for *V. vulgaris* as they had the same Kappa score, CV error was used to select from the two equivalent Kappa score models. P and DR indicate the predictor data and dimension reduction method used along with the models selected as best or worst models.

Comparison between TD-DR-MT combinations (Table S4.1) showed that model type could make a difference in prediction accuracy for some presence data, especially when the presence/pseudo-absence data were less reliable. For example, there were no statistically significant differences between Kappa scores for LOGR, QDA, CART, and SVM predictions for *V. vulgaris*. On the other hand, there was a statistically significant difference between Kappa scores of LOGR/QDA and SVM/CART for *T. pityocampa*, where the machine learning methods handled the low presence data prevalence better.

3.6. Species Distribution Predictions and Uncertainty

Model predictions were not examined until all the model performance score analyses were finalized. Once the best and worst modeling component combinations for all species were identified (Table 4), the corresponding predictions were examined. Most predictions were from the top five best models identified areas, which were well described as native or introduced geographical ranges of the five species studied (Note S5).

However, in all cases, the kappa scores of the best P-DR-MT combinations selected for the five species were not significantly different from the second best combinations. In some cases, the difference between the Kappa scores was not significant for the first 5–6 combinations. All of the worst modeling component combinations for the five species, however, had significantly lower kappa scores from the second worst modeling component combinations for the respective species. The second best modeling component combination (Figure 6C) for *A. gracilipes*, with the same dataset but a different dimension reduction method from the best combination (Figure 6B) with the second highest Kappa value, is shown in Figure 6. Clearly, the prediction of the second best Kappa model was similar to the best Kappa model. Assessing the probabilistic predictions for the five species showed that three models with the highest kappa scores, may have over-predicted the geographic ranges of the species under study (Figure S6). No additional validation data were available, thus assessment was done visually and only extreme predicted areas known to be outside the biological tolerance of the species were considered as over-predictions.

The availability of various distribution prediction scenarios based on the different modeling component combinations allowed for the generation of standard deviation maps that could be used as uncertainty measures for predictions. The multi-scenario mean prediction and the associated prediction standard deviation maps for *A. albopictus* are given in Figure 7A,B, respectively (Figure S7 shows data for the other species).

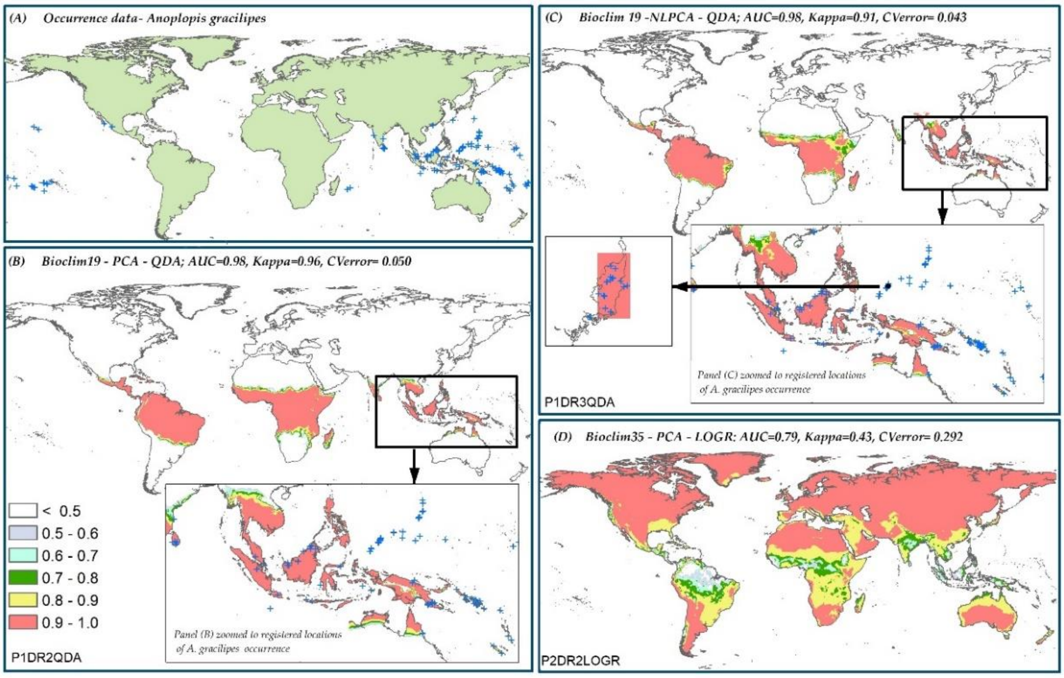
(A) Occurrence data- Anoplopis gracilipes
(C) Bioclim 19 -NLPCA - QDA; AUC=0.98, Kappa=0.91, CVerror= 0.043
(B) Bioclim19 - PCA - QDA; AUC=0.98, Kappa=0.96, CVerror= 0.050
Panel (C) zoomed to registered locations of A. gracilipes occurrence
P1DR3QDA
(D) Bioclim35 - PCA - LOGR: AUC=0.79, Kappa=0.43, CVerror= 0.292
< 0.5
0.5 - 0.6
0.6 - 0.7
0.7 - 0.8
0.8 - 0.9
0.9 - 1.0
Panel (B) zoomed to registered locations of A. gracilipes occurrence
P1DR2QDA
P2DR2LOGR

Figure 6. Predicted probability of presence for *A. gracilipes*. (**A**) Occurrence data, (**B**) the best model combination for *A. gracilipes*, (**C**) the second best model combination for *A. gracilipes*, and (**D**) the worst model combination for *A. gracilipes*.

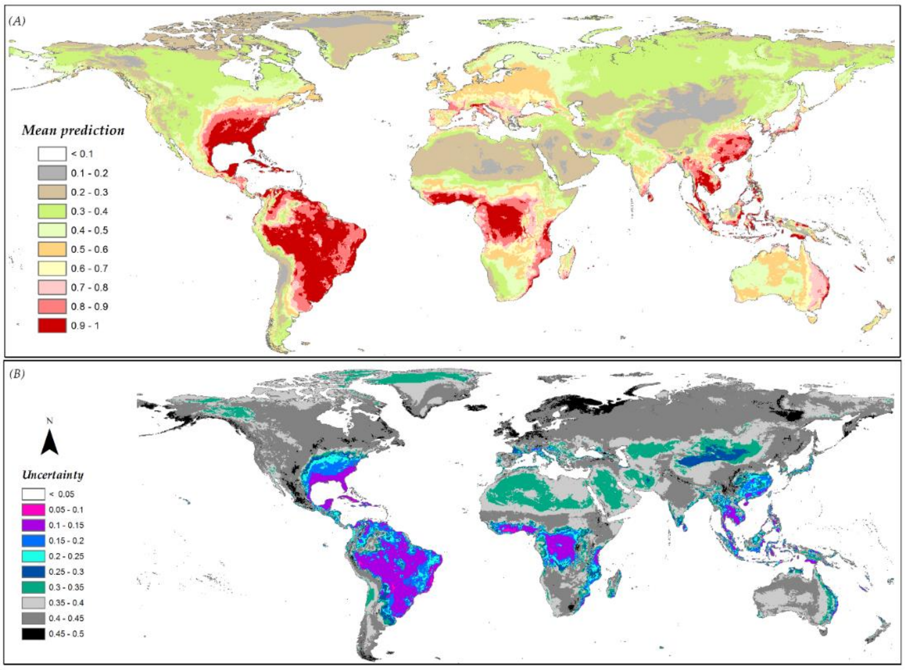

Figure 7. (**A**) Mean predicted presence across all scenarios for *A. albopictus*; (**B**) the associated uncertainty around the mean prediction within the multi-scenario modeling framework. Grey shades show higher uncertainty whereas purplish-bluish shades show lower uncertainty in the form of low SD among replicates.

The probability density for the standard deviation of model predictions given by the different modeling components is shown in Figure 8D for *A. albopictus* and Figure S8 for the other species.

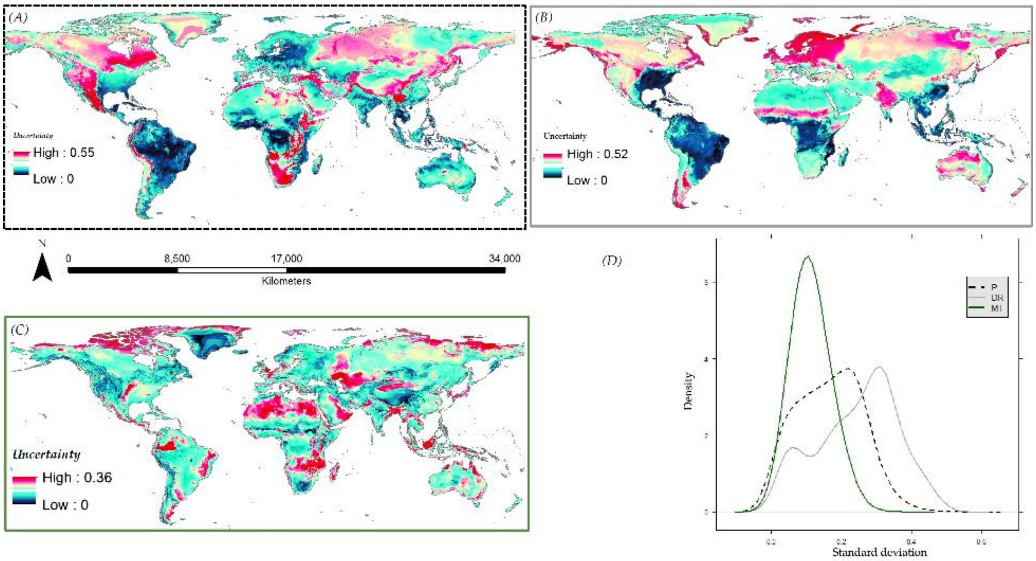

Figure 8. Spatial pattern of variability according to, (**A**) predictor data—P, (**B**) dimension reduction—DR, (**C**) model type—MT, and (**D**) the probability density of predicted presences according to the P, DR, and MT for *A. albopictus*.

4. Discussion

According to the MANOVA results, the effects of major modeling components such as model type, dimension-reduction, species dataset, and predictor dataset on model performance are in accordance with previous studies that investigated sources of uncertainty in SDM predictions [5,9,10,66–68]. Similarly, the per-species univariate analysis of variance conformed to the general trend with regard to the modeling components order of significance except for one case. For *A. gracilipes*, only predictor data had a significant effect on model performance, unlike the other four species where model type had the largest effect. This specific case shows that it is possible to have exceptions to the established trend requiring a case-by-case investigation to identify the most important modeling component whenever the modeling scenario changes. Change in modeling scenario here refers to change in occurrence data, species, predictor data, models, and any data pre-processing methods used. In this study, the *A. gracilipes* presence locations clearly clustered in the environmental feature space. Species with a limited environmental niche can be mapped using less complex models and limited variables, requiring a less complex modeling scenario (Reference [19], and references therein).

High presence data prevalence was associated with high model Kappa scores in all cases, however, prediction accuracy does not necessarily follow the size of the presence dataset as reported by Elith, et al. [5], based on their factorial study involving 226 species and 17 SDMs. Therefore, we drew no definitive conclusion from the strong correlation between presence data prevalence and high Kappa scores.

Predictor datasets: According to the frequency analysis for selected variables, more individual variables common to P1, P2, and P3 datasets were consistently included in the models than variables unique to P2 and P3 datasets (ranked 9th or higher out of the total rank of 18). The second most included set of variables were unique to dataset P2 and P3 (ranked 9th or higher out of the total rank of 18), therefore it is recommended to use the P2 dataset [31], which includes all the top half of the ranked list of predictors unless the modeler has good evidence that the target species distribution can be adequately described by temperature and precipitation derived variables alone (Table S2). Elevation was the only variable unique to the P3 dataset that was consistently selected. The other three variables: slope, aspect, and hill-shade were only included in three models. Therefore, it seems that these three topographical variables may be more useful for higher resolution data at local scales than global or regional scale studies where elevation data can be used as a proxy for those variables.

Dimension reduction method: Nonlinear principal component analysis (NLPCA) appeared to perform well based on comparisons of Kappa and cross-validation error. The most important information concerning dimension reduction methods, however, was obtained from assessing the actual predictions rather than the statistical pairwise comparisons of model Kappa scores.

Three of the five models with the highest Kappa score selected for the five species had used the h-NLPCA (DR3) as a dimension reduction method, and when their corresponding predictions were assessed it was apparent that all three models had over-predicted. Deciding whether a model has over-predicted is not an easy task unless there is additional external validation data that covers areas beyond what is covered by the training/test data. In the case of the three models selected for *A. albopictus*, *D. v. virgifera*, and *V. vulgaris*, determining whether they were over-predicting was straightforward because the areas that were incorrectly predicted were in areas where the environmental conditions were outside the known biological tolerance of these three species (Figure S6). For example, some of these areas were Greenland for *A. albopictus*, Sahara, and the Middle East for *D. v. virgifera* and *V. vulgaris*.

Exploration of relative locations for the presence points of the three species and their corresponding pseudo-absences points selected from the h-NLPCA (DR3) transformed data, revealed a possible reason for the over-prediction. The pseudo-absences selected from the h-NLPCA transformed datasets were highly discriminated from the presence points, and highly localized in the predictor feature space. While high discrimination between presences and pseudo-absences in the feature space can be desirable for a clear characterization of suitable and unsuitable habitats, the highly localized pseudo-absence points are problematic, as that means less information regarding what constitutes unsuitable conditions in the environmental space is available. The high Kappa scores for the h-NLPCA dimension reduction method indicate that h-NLPCA may be a useful tool in species distribution modeling, provided that the tendency for models to over-fit on h-NLPCA-transformed data can be addressed through incurring a training gain penalty [69] or a regularization scheme [70].

Model type: Machine-learning methods were consistently highly ranked in performance for species with high data prevalence covering large portions of environmental space, while QDA performed better with species that had low prevalence—where occurrences occupied a localized area in environmental space. Similar results were reported by Segurado and Araújo [71], in their study that evaluated commonly used species distribution models. This result leads to the conclusion that each species should be treated individually, and that model selection should be solely based on the occurrence data used and not on recommendations from other studies that have used very different presence data and environmental variables. Several caveats should be noted. Some models maybe positively or negatively affected by the pseudo-absences to presences ratio during model training. For example, regression-modeling techniques perform better when disproportionately more pseudo-absences (zeros) are used for fitting [72–74]. In our study, we kept the number of pseudo-absences equal to presences, to ensure that factors not measured for effect are kept constant for an unbiased comparison of model outcomes. In addition, we believe that the LOGR model may, on average, have performed worse than other models because different regularization options were not tested [73]. Our objective was to test the models in their commonly used formats, while individually parameterizing each model was outside of the scope of this study.

Performance measures: Indices based on the confusion matrix are the most used methods for model performance measurement in species distribution modeling [51]. These methods have been widely and successfully used in other disciplines, especially in clinical studies, long before they were adapted for ecological modeling [75]. However, methods based on the confusion matrix are not always sufficient for model validation in the ecological context, in cases where minimal training data is used for species distribution predictions that cover a much greater geographical area than the training data [75–77]. This training data/prediction imbalance is especially pronounced when SDMs are used for regional or global studies. An example in this study is that none of the models with the highest kappa scores were statistically different from the second or third highest score models. This equivalence effectively means that different modeling factor combinations used in the second or third highest kappa score models could give similar or sometimes better (in case of models that over-fit or extrapolate [78,79]) predictions than the actual highest Kappa score model. There is therefore a need to devise additional model prediction evaluation tools beyond those generated from a confusion matrix.

5. Conclusions

This study shows that the predictive performance of different model types depends mainly on data pre-processing, in other words, on pseudo-absence dataset development, dimension-reduction method, species- and predictor-dataset selection. Thus, performance comparisons among different model types cannot be applied unless all data pre-processing factors are kept constant. The model-type comparisons applied for the same species and predictor datasets and under the same dimension-reduction method showed that machine learning models (mainly SVM) outperform other model types, with the exception that less complex statistical models can perform well for species with a limited environmental niche. h-NLPCA was also a highly ranked dimension-reduction method. Variation in SDM predictions does not always necessarily occur due to reasons inherent in the model types used. Thus, it is possible to achieve better SDM predictions using appropriate modeling components such as predictors and dimension reduction methods that fit the available data. The variation in species distribution prediction was observed when we used different modeling components coupled with the low discriminatory power of evaluation methods among top performing models, supports two recommendations for modeling species distributions to increase prediction certainty: 1) where possible explore the use of multi-scenario modeling frameworks where a variation of environmental variables, dimension-reduction methods, and model types are tested in a standardized manner before picking the best modeling combination; and 2) use a second model performance measure to allow discriminating model combinations that might have identical or very similar scores based on the first model performance measure. Performing species distribution predictions in a multi-scenario modeling framework can also provide some measure of uncertainty through assessment of model consensus between predictions made for the same species, using different modeling components. When resources available for control or eradication of an invading insect population are limited, the design of sampling protocols for detection may be focused on predicted locations that are accompanied by high model consensus or low-uncertainty hotspots.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-4450/10/3/65/s1, S1: Canonical analyses results; Table S2: Ranks of variables based on proportion of inclusion in models; Figure S3: External validation data for *V. vulgaris* in New Zealand; Table S4: Comparison between different modeling component combinations; Note S5: Distribution predictions for the five species in this study; Figure S6: Species distribution maps according to three different modeling component combinations; Figure S7: Ensemble mean predictions and uncertainty maps; Figure S8: Map of uncertainty by modeling components.

Author Contributions: Conceptualization, S.D.S. and S.P.W.; methodology, S.D.S. and S.P.W.; software, S.D.S. and S.P.W.; validation, S.D.S. and S.P.W.; formal analysis, S.D.S.; investigation, S.D.S. and S.P.W.; resources, S.P.W.; data curation, S.D.S.; writing—original draft preparation, S.D.S.; writing—review and editing, S.D.S. and S.P.W.; visualization, S.D.S.; supervision, S.P.W.; project administration, S.P.W.; funding acquisition, S.P.W.

Funding: This research was funded by the Bio-Protection Research Centre (BPRC); https://bioprotection.org.nz/.

Acknowledgments: The authors wish to thank PRATIQUE (European Commission's 7th Framework Programme) for the Western corn rootworm data, and GBIF for the other species data used in this study. The authors would also like to thank Carsten F. Dormann and Steven Higgins for their valuable comments on the earlier versions of this manuscript, and Gwenaël Leday and Tak Ikeda for the development of the original multimodel framework software. We are very grateful for the time and effort the two anonymous reviewers spent on our article, their input immensely improved this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in writing of manuscript, or in the decision to publish the results.

Appendix A. Background on h-NLPCA Dimension Reduction Method

The description of the h-NLPCA by Scholz and Vigario [37] is adapted here to give a brief background. The h-NLPCA (Figure A1C) is built upon a pre-existing form of a multilayer perceptron [80] network (Figure A1A) with an auto-associative neural network topology also known as bottleneck or hourglass topology [81].

The auto-associative network is a linear multilayer perceptron auto-encoder that has the same number of inputs (nodes) as outputs, and a hidden layer with fewer nodes. The weights in the network change while learning to minimize the mean square error. Due to this bottleneck architecture, the equivalent number of inputs and outputs, and the algorithm that minimizes the mean square error, it was possible to converge the *n* features to the *n*th dimension in the linear PCA feature space for a given *n x m* matrix [82].

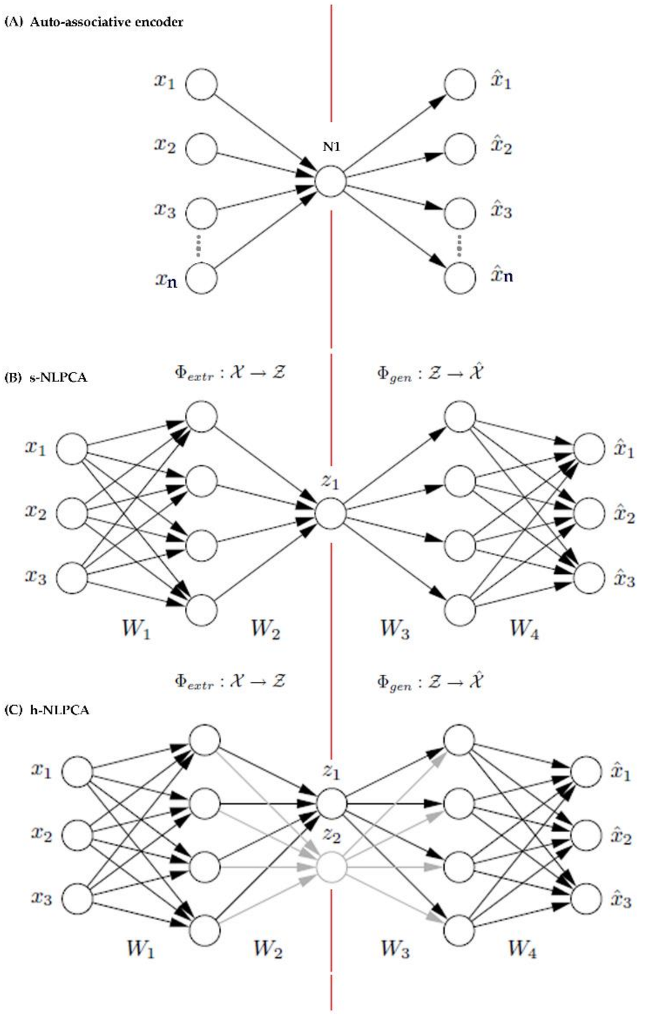
(A) Auto-associative encoder
x_1
x_2
x_3
x_n
N1
$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_3$
$\hat{x}_n$
(B) s-NLPCA
$\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$
$\Phi_{gen} : \mathcal{Z} \rightarrow \hat{\mathcal{X}}$
x_1
x_2
x_3
z_1
$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_3$
W_1
W_2
W_3
W_4
$\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$
$\Phi_{gen} : \mathcal{Z} \rightarrow \hat{\mathcal{X}}$
(C) h-NLPCA
x_1
x_2
x_3
z_1
z_2
$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_3$
W_1
W_2
W_3
W_4

Figure A1. Topologies of the linear auto-encoder (**A**) the auto-encoder can have more than one node as long as the number of nodes at the hidden layer are fewer than the nodes at the input and output, which have equal number of nodes [81], the s-NLPCA (B) and the h-NLPCA (C) auto-associative neural networks. The illustrations for the topologies of (**B**) s-NLPCA [3-4-4-4-3], and (**C**) h-NLPCA ([3-4-2-4-3] + [3-4-1-4-3]) networks were taken from Scholz, et al. [83]. The illustration for (**A**) the auto-encoder [4-1-4] was adapted from Marivate, et al. [81]. The numbers in brackets indicate the network topology description that gives the number of nodes in the input, any hidden layers and output layers in that order.

Kramer [84] then expanded the above auto-associative encoder into a non-linear PCA by adding two layers of nodes with non-linear functions at the start and end of the auto-associative encoder (Figure A1B). The extension enabled the linear auto-associative network to extract the principal components from non-linear feature subspace.

In a nutshell, what Scholz and Vigario [37] have done was to extend the s-NLPCA into a hierarchical auto-associative neural network (non-linear PCA encoder) by superimposing an extra non-linear network with a function that applies hierarchy constraints to the feature space in the same way as the linear PCA does [37]. This gave rise to the h-NLPCA (Figure A1C). The symmetrical NLPCA (s-NLPCA) was similar to h-NLPCA by mapping data onto a non-linear feature sub-space, but it lacks the ability to discriminate features. Generally, a s-NLPCA (Figure A1B) is sufficient if the problem involves only reducing dimensions and does not require feature selection [37,38]. However, in the context of this study, a method that does dimension reduction and is capable of identifying features in the non-linear feature space, is preferable. Since there is no subsequent feature selection step specified for datasets that are treated with dimension reduction, h-NLPCA is the appropriate choice because the hierarchical learning algorithm allows for feature selection as well as dimension reduction.

Both s-NLPCA and h-NLPCA are extensions of the linear auto-encoder. In both cases, the left side of the red line in the middle shows the first part of the dimension reduction process where data are extracted non-linearly from the inputs in [x1, x2, x3 . . . ..], and linearly decoded at [z_1, (z_2)] (function Φ_{extr}). The right side of the red line shows where data are linearly decoded from [z_1, (z_2)], and are non-linearly generated at the output [x1, x2, x3 . . .] (function Φ_{gen}). For h-NLPCA, an additional 3-4-1-4-3 topology network is transposed on top of the s-NLPCA topology so that learning error is separately computed, 1) for the sub-network (E_1), and 2) on the sub-network + the whole network ($E_{1,2}$). Learning error is later added to produce the total hierarchic error ($E = E_1 + E_{1,2}$), which is used to update the weights through the whole network. This hierarchic learning enables the h-NLPCA to do feature extractions as well as dimension reductions. Refer Scholz and Vigario [37] and Scholz, et al. [83] for detailed model specification. The above description was also modified from the same references.

References

1. Elith, J.; Leathwick, J.R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu. Rev. Ecol. Evol. Syst.* **2009**, *40*, 677–697. [CrossRef]
2. Elith, J.; Burgman, M.A.; Regan, H.M. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Model.* **2002**, *157*, 313–329. [CrossRef]
3. Thuiller, W. Patterns and uncertainties of species' range shifts under climate change. *Glob. Change Biol.* **2004**, *10*, 2020–2027. [CrossRef]
4. Araújo, M.B.; Guisan, A. Five (or so) challenges for species distribution modelling. *J. Biogeogr.* **2006**, *33*, 1677–1688. [CrossRef]
5. Elith, J.; Graham, C.H.; Anderson, R.P.; Dudik, M.; Ferrier, S.; Guisan, A.; Hijmans, R.J.; Huettmann, F.; Leathwick, J.R.; Lehmann, A.; et al. Novel methods improve prediction of species; distributions from occurrence data. *Ecography* **2006**, *29*, 129–151. [CrossRef]
6. Hartley, S.; Lester, P.J.; Harris, R. Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecol. Lett.* **2006**, *9*, 1068–1079. [CrossRef] [PubMed]
7. Pearson, R.G.; Thuiller, W.; Araújo, M.B.; Martinez-Meyer, E.; Brotons, L.; McClean, C.; Miles, L.; Segurado, P.; Dawson, T.P.; Lees, D.C.; et al. Model-based uncertainty in species range prediction. *J. Biogeogr.* **2006**, *33*, 1704–1711. [CrossRef]
8. Araújo, M.B.; New, M. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **2007**, *22*, 42–47. [CrossRef] [PubMed]
9. Dormann, C.F.; Purschke, O.; Marquez, J.R.G.; Lautenbach, S.; Schroder, B. Components of uncertainty in species distribution analysis: A case study of the Great Grey Shrike. *Ecology* **2008**, *89*, 3371–3386. [CrossRef] [PubMed]
10. Buisson, L.; Thuiller, W.; Casajus, N.; Lek, S.; Grenouillet, G. Uncertainty in ensemble forecasting of species distribution. *Glob. Change Biol.* **2010**, *16*, 1145–1157. [CrossRef]

11. Venette, R.C.; Kriticos, D.J.; Magarey, R.D.; Koch, F.H.; Baker, R.H.A.; Worner, S.P.; Raboteaux, N.N.G.; McKenney, D.W.; Dobesberger, E.J.; Yemshanov, D.; et al. Pest Risk Maps for Invasive Alien Species: A Roadmap for Improvement. *BioScience* **2010**, *60*, 349–362. [CrossRef]
12. De Marco, P.J.; Nóbrega, C.C. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS ONE* **2018**, *13*, e0202403. [CrossRef] [PubMed]
13. Wang, G.; Gertner, G.Z.; Fang, S.; Anderson, A.B. A Methodology for Spatial Uncertainty Analysis Of Remote Sensing and GIS Products. *Photogram. Eng. Rem. Sens.* **2005**, *71*, 1423–1432. [CrossRef]
14. Yemshanov, D.; Koch, F.H.; Lyons, D.B.; Ducey, M.; Koehler, K. A dominance-based approach to map risks of ecological invasions in the presence of severe uncertainty. *Divers. Distrib.* **2011**, *18*, 33–46. [CrossRef]
15. Busby, J.R.; McMahon, J.P.; Hutchinson, M.F.; Nix, H.A.; Ord, K.D. BIOCLIM—A bioclimate analysis and prediction system. *Plant Prot. Q.* **1991**, *6*, 8–9.
16. Carpenter, G.; Gillison, A.N.; Winter, J. DOMAIN: A flexible modelling procedure for mapping potential distributions of plants and animals. *Biodivers. Conserv.* **1993**, *2*, 667–680. [CrossRef]
17. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, NY, USA, 1995.
18. Tsoar, A.; Allouche, O.; Steinitz, O.; Rotem, D.; Kadmon, R. A comparative evaluation of presence-only methods for modeling species distribution. *Divers. Distrib.* **2007**, *13*, 397–405. [CrossRef]
19. Jiménez-Valverde, A.; Lobo, J.M.; Hortal, J. Not as good as they seem: the importance of concepts in species distribution modeling. *Divers. Distrib.* **2008**, *14*, 885–890. [CrossRef]
20. Chefaoui, R.M.; Lobo, J.M. Assessing the effects of Pseudo-absence on predictive distribution model performance. *Ecol. Model.* **2008**, *210*, 478–486. [CrossRef]
21. Senay, S.D.; Worner, S.P.; Ikeda, T. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modeling. *PLoS ONE* **2013**, *8*, e71218. [CrossRef] [PubMed]
22. Kearney, M.; Porter, W. Mechanistic niche modeling: combining physiological and spatial data to predict species' ranges. *Ecol. Lett.* **2009**, *12*, 334. [CrossRef] [PubMed]
23. Pereira, J.M.C.; Itami, R.M. GIS-based habitat modeling using logistic multiple regression: A study of the Mt. Graham red squirrel. *Photogramm. Eng. Remote Sens.* **1991**, *57*, 1476–1482.
24. Zimmermann, N.E.; Edwards, T.C.; Moisen, G.G.; Frescino, T.S.; Blackard, J.A. Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *J. Appl. Ecol.* **2007**, *44*, 1058–1060. [CrossRef] [PubMed]
25. Austin, M.P.; Van Niel, K.P. Improving species distribution models for climate change studies: variable selection and scale. *J. Biogeogr.* **2010**, *38*, 1–8. [CrossRef]
26. Heikkinen, R.K.; Luoto, M.; Kuussaari, M.; Pöyry, J. New insights into butterfly–environment relationships using partitioning methods. *Proc. R. Soc. B* **2005**, *272*, 2203–2210. [CrossRef] [PubMed]
27. Luoto, M.; Virkkala, R.; Heikkinen, R.K.; Rainio, K. Predicting bird species richness using remote sensing in boreal agricultural-forest mosaics. *Ecol. Appl.* **2004**, *14*, 1946–1962. [CrossRef]
28. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
29. Hijmans, R.J.; Cameron, S.; Parra, J. WORLDCLIM. Available online: http://www.worldclim.org/ (accessed on 15 June 2018).
30. Hijmans, R.J.; Cameron, S.E.; Parra, J.L. BIOCLIM. Available online: http://www.worldclim.org/bioclim (accessed on 15 June 2018).
31. Kriticos, D.J.; Webber, B.L.; Leriche, A.; Ota, N.; Macadam, I.; Bathols, J.; Scott, J.K. CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods Ecol. Evol.* **2011**, *3*, 53–64. [CrossRef]
32. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **2005**, *25*, 1965–1978. [CrossRef]
33. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
34. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
35. Dupin, M.; Reynaud, P.; Jarošík, V.; Baker, R.; Brunel, S.; Eyre, D.; Pergl, J.; Makowski, D. Effects of the Training Dataset Characteristics on the Performance of Nine Species Distribution Models: Application to Diabrotica virgifera virgifera. *PLoS ONE* **2011**, *6*, e20957. [CrossRef] [PubMed]

36. Hirzel, A.H.; Hausser, J.; Chessel, D.; Perrin, N. Ecological-Niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* **2002**, *83*, 2027. [CrossRef]
37. Scholz, M.; Vigario, R. Nonlinear PCA: A new hierarchical approach. In Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 24–26 April 2002; pp. 439–444.
38. Gorban, A.N.; Zinovyev, A.Y. *Elastic Maps and Nets for Approximating Principal Manifolds and Their Application to Microarray Data Visualization*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 96–130.
39. Iturbide, M.; Bedia, J.; Herrera, S.; Del Hierro, O.; Pinto, M.; Gutiérrez, J.M. A framework for species distribution modelling with improved pseudo-absence generation. *Ecol. Model.* **2015**, *312*, 166–174. [CrossRef]
40. Kampichler, C.; Wieland, R.; Calmé, S.; Weissenberger, H.; Arriaga-Weiss, S. Classification in conservation biology: A comparison of five machine-learning methods. *Ecol. Inform.* **2010**, *5*, 441–450. [CrossRef]
41. Worner, S.P.; Gevrey, M.; Ikeda, T.; Leday, G.; Pitt, J.; Schliebs, S.; Soltic, S. Ecological Informatics for the Prediction and Management of Invasive Species. In *Springer Handbook of Bio-/Neuroinformatics*; Springer Nature: New York, NY, USA, 2014; pp. 565–583.
42. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available online: http://www.R-project.org/ (accessed on 29 October 2012).
43. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics With S*, 4th ed.; Springer: New York, NY, USA, 2002.
44. Lorena, A.C.; Jacintho, L.F.; Siqueira, M.F.; De Giovanni, R.; Lohmann, L.G.; De Carvalho, A.C.; Yamamoto, M. Comparing machine learning classifiers in potential distribution modelling. *Expert Syst. Appl.* **2011**, *38*, 5268–5275. [CrossRef]
45. Garzón, M.B.; Blažek, R.; Neteler, M.; De Dios, R.S.; Ollero, H.S.; Furlanello, C. Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecol. Model.* **2006**, *197*, 383–393.
46. Way, M.J.; Scargle, J.D.; Ali, K.M.; Srivastava, A.N. *Advances in Machine Learning and Data Mining for Astronomy*; Taylor & Francis: Abington, UK, 2012.
47. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
48. Allouche, O.; Tsoar, A.; Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **2006**, *43*, 1223–1232. [CrossRef]
49. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
50. Jiménez-Valverde, A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Glob. Ecol. Biogeogr.* **2011**, *21*, 498–507. [CrossRef]
51. Fielding, A.H.; Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **1997**, *24*, 38–49. [CrossRef]
52. Worner, S.P.; Ikeda, T.; Leday, G.; Joy, M. *Surveillance Tools for Freshwater Invertebrates*; Biosecurity Technical Paper 2010/21; Ministry Agriculture Forestry NZ: Wellington, New Zealand, 2010.
53. Mahalanobis, P.C. On the generalized distance in statistics. *J. Asiat. Soc. Bengal* **1930**, *26*, 541–588.
54. Boeschen, L.E.; Koss, M.P.; Figueredo, A.J.; Coan, J.A. Experiential avoidance and post-traumatic stress disorder: A cognitive mediational model of rape recovery. *J. Aggress. Maltreatment Trauma* **2001**, *4*, 211–245. [CrossRef]
55. Box, G.E. A general distribution theory for a class of likelihood criteria. *Biometrika* **1949**, *36*, 317–346. [CrossRef] [PubMed]
56. Howell, D. Statistical methods for psychology Thomson Wadsworth. *Belmont CA* **2007**, 1–739.
57. De Mendiburu, F. Agricolae: Statistical Procedures for Agricultural Research R Package Version 1.1-2. Available online: http://CRAN.R-project.org/package=agricolae (accessed on 12 September 2012).
58. Friendly, M.; Fox, J. *Candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis*, R package 0.6-5. 2013.
59. González, I.; Déjean, S. *CCA: Canonical Correlation Analysis*, R package 1.2. 2012.
60. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2009.
61. Fox, J.; Friendly, M.; Monette, G. *Heplots: Visualizing Tests in Multivariate Linear Models*, R package 1.0-11. 2013.
62. Walsh, C.; Nally, R.M. *hier.part: Hierarchical Partitioning*, R package 1.0-4. 2013.

63. Hothorn, T.; Bretz, F.; Westfall, P. Simultaneous inference in general parametric models. *Biom. J.* **2008**, *50*, 346–363. [CrossRef] [PubMed]
64. Chevan, A.; Sutherland, M. Hierarchical partitioning. *Am. Stat.* **1991**, *45*, 90–96.
65. MacNally, R. Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodivers. Conserv.* **2000**, *9*, 655–671. [CrossRef]
66. Lawler, J.J.; White, D.; Neilson, R.P.; Blaustein, A.R. Predicting climate-induced range shifts: model differences and model reliability. *Glob. Change Biol.* **2006**, *12*, 1568–1584. [CrossRef]
67. Diniz-Filho, J.A.F.; Mauricio Bini, L.; Fernando Rangel, T.; Loyola, R.D.; Hof, C.; Nogués-Bravo, D.; Araújo, M.B. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography* **2009**, *32*, 897–906. [CrossRef]
68. Roura-Pascual, N.; Brotons, L.; Peterson, A.T.; Thuiller, W. Consensual predictions of potential distributional areas for invasive species: A case study of Argentine ants in the Iberian Peninsula. *Biol. Invasions* **2008**, *11*, 1017–1031. [CrossRef]
69. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
70. Jiménez, A.A.; García Márquez, F.P.; Moraleda, V.B.; Gómez Muñoz, C.Q. Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis. *Renew. Energy* **2019**, *132*, 1034–1048. [CrossRef]
71. Segurado, P.; Araújo, M.B. An evaluation of methods for modelling species distributions. *J. Biogeogr.* **2004**, *31*, 1555–1568. [CrossRef]
72. Barbet-Massin, M.; Jiguet, F.; Albert, C.H.; Thuiller, W.; Barbet-Massin, M. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* **2012**, *3*, 327–338. [CrossRef]
73. Gastón, A.; García-Viñas, J.I. Modeling species distributions with penalised logistic regressions: A comparison with maximum entropy models. *Ecol. Model.* **2011**, *222*, 2037–2041. [CrossRef]
74. Wisz, M.; Guisan, A. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecol.* **2009**, *9*, 8. [CrossRef] [PubMed]
75. McPherson, J.M.; Jetz, W.; Rogers, D.J. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.* **2004**, *41*, 811–823. [CrossRef]
76. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E.R. Small-sample precision of ROC-related estimates. *Bioinformatics* **2010**, *26*, 822–830. [CrossRef] [PubMed]
77. Lobo, J.M. More complex distribution models or more representative data? *Biodiv. Inf.* **2008**, *5*, 14–19. [CrossRef]
78. Elith, J.; Simpson, J.; Hirsch, M.; Burgman, M.A. Taxonomic uncertainty and decision making for biosecurity: spatial models for myrtle/guava rust. *Australas. Plant Pathol.* **2012**, *42*, 43–51. [CrossRef]
79. Raes, N.; Aguirre-Gutiérrez, J. Modeling Framework to Estimate and Project Species Distributions Space and Time. *Mt. Clim. Biodivers.* **2018**, *309*.
80. Bishop, C.M. *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, NY, USA, 1995.
81. Marivate, V.N.; Nelwamodo, F.V.; Marwala, T. Autoencoder, principal component analysis and support vector regression for data imputation. *arXiv preprint*, 2007; arXiv:0709.2506.
82. Baldi, P.; Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw.* **1989**, *2*, 53–58. [CrossRef]
83. Scholz, M.; Fraunholz, M.; Selbig, J. Nonlinear Principal Component Analysis: Neural Network Models and Applications. In *Principal Manifolds for Data Visualization and Dimension Reduction*; Gorban, A.N., Ed.; Springer: New York, NY, USA, 2008; pp. 44–67.
84. Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **1991**, *37*, 233–243. [CrossRef]