Chapter 2

The relative importance of factors influencing invasive water hyacinth occurrence: An Earth Observation and explainable machine learning method.

**Abstract**

Variable success in the control of water hyacinth prompted an investigation into the ecological and socio-economic contexts that influence the occurrence of the weed. To assist the selection of management strategies and risk-prioritisation of sites, through a data-driven and evidence-based approach. The relationship between water hyacinth occurrence and its likely drivers were identified and their relative importance investigated in a spatially explicit manner. This study demonstrates the value of recent advances in explainable Artificial Intelligence (xAI), Earth Observation (EO) and cloud computing. This low-cost computer-based method used SHaply Additive exPlanations (SHAP) to identify and explain species-environment interactions learnt by species distribution models. Many of the results agree with known physiological constraints limiting water hyacinth growth but were also able to provide novel insights attributed to the combination of EO data, and xAI. These may be valuable to inform criteria for an early warning system, or stimulate future lab and field-based validation experiments that can concurrently assist in preventing the spread of water hyacinth.

**Keywords:** satellite, remote sensing, Google Earth Engine, interpretable artificial intelligence, habitat suitability.

1. **Introduction**

Across tropical, sub-tropical, and warm-temperate regions [1,2], water hyacinth (*Pontederia* (previously *Eichhornia*) *crassipes* Mart. (Pontederiaceae) [3]) has frequently been linked to perturbed aquatic environments constraining ecosystem service provision [4,5]. Water hyacinth has extended its non-native range into South Africa, were it was first is one of the countries with a long history of water hyacinth infestation is, water hyacinth was first recorded during 1908. Since then there has been a large investment into the development of IAAP control strategies (for example, [6–8]). However, water hyacinth persists and was estimated to cover 417.74 square kilometres of South Africa during 2013 [9]. This is partly attributed to current reactive management strategies [10], along with the slow adoption of recent technological advances in earth observation, cloud computing and machine learning. This results in a lack of actionable insights from early IAAP incursion warnings and prevents timely responses to new incursions across large extents of the plants’ introduced range. This is despite a higher chance of successful management when recently introduced alien populations are targeted [11].

The development of a pre-emptive management strategy relies on understanding the ecological and socio-economic context that limits or promotes water hyacinth establishment and spread [12]. For the effective use of limited resources, this knowledge should be available in a cost-effective manner, at local to national scales. In this way, it can be integrated into decision making processes that underpin the pre-selection and prioritisation of different control strategies, on a site-by-site basis [13,14].

IAAP management strategies centered around the early response to incursions into new water systems can benefit [15–20] from Earth Observation (EO) for early detection (through surveillance), and early warnings (through modelling). While monitoring may provide an indication of new incursions and the actual IAAP distribution [15], modelling habitat susceptibility and habitat suitability for an IAAP may provide an indication of probable future incursions and areas that should be closely monitored. Habitat suitability models or Species Distribution Models (SDMs) represent the relative likelihood of an IAAP establishment should the species be introduced or disperse to each location in the modelled landscape [21] and are well established [22,23]. This contrasts with the methods for modelling the risk of introduction and spread.

While a comprehensive Risk Analysis for Alien Taxon (RAAT) has been established [24], it is largely qualitative in nature and suited to coarse and broad extent analysis. Previous studies have investigated potential water hyacinth distribution under varying climate change scenarios using mechanistic climate matching models (CLIMEX) [25] and correlative SDMs [25,26]. However, these studies were limited by the number of environmental variables considered, and did not consider the spatial variability of the species-environment relationships learnt by the models (limited to SDMs) during interpretation. Moreover, numerous previous SDM studies did not consider recommended modelling guidelines during their analysis [27]. This study therefore seeks to harness recent advances in explainable Artificial Intelligence (xAI) while adhering to recommended SDM modelling guidelines for the identification, exploration, and understanding of spatially variable species-environment relationships captured by a SDM [27,28].

EO coupled with recent advances in xAI are suited to multi-scale spatio-temporal SDM analysis (for example, [29]) owing to 1) the recent availability of free cloud computing infrastructure that has lowered the barrier for large-scale EO data analysis, 2) the public release of fine-scale but large-extent EO products that benefit from free data policies along with global, systematic, frequent acquisition of EO data (for example, [30–32]), and open-source libraries [33]. Together, these factors have enabled the production of numerous EO-derived data products that represent ecological, hydrological, climatological, social, and topographical phenomenon across the earth’s surface [34,35]. The combined input of these derived variables with species occurrences into ML algorithms have become popular choices for determining a species’ environmental range via SDMs [35]. This can be attributed to their ability to model non-linear species-environment responses to environmental gradients and their predictive performance, despite high dimensionality [36,37].

To identify and explore the 1) relative variable importance of likely water hyacinth drivers and, 2) SDM-captured species-environment responses to environmental gradients, the recent SHaply Additive exPlanations (SHAP) is used [29]. Because of its theoretical justification and analytical gains of over other proposed xAI tools (for example, Local Interpretable Model-agnostic Explanations (LIME) and Mean Decrease in Impurity (MDI)), SHAP is preferred and increasingly being adopted for model interpretation [28,38]. The model agnostic SHAP stem from Shapley values, a concept coming from game theory to determine the contribution of each player (an environmental predictor of water hyacinth) in a coalition towards the outcome of the game (reproducing the outcome of the model- probability of water hyacinth occurrence). Here, each variable’s (players’) contribution to the prediction (reward) is evaluated [39]. A limited number of studies have demonstrated the usefulness of SHAP in offering novel insights that deepen understanding, or reveal new scientific discoveries across a variety of domains [40], including ecology (for example, [29,41,42]).

This study demonstrates the utility of SHAP, EO-derived data products and cloud computing made available through Google Earth Engine (GEE) and Google Collaboratory to; 1) determine the relative overall importance of biotic and abiotic factors that influence the occurrence of water hyacinth for a single waterbody or across waterbodies within a province, or across South Africa, 2) capture the relationship between socio-economic and environmental factors and the probability of water hyacinth occurrence, and 3) elucidate the influence of interaction effects between variables on the probability of water hyacinth occurrence.

1. **Methods**

The following sections describe each step of the workflow in detail (Figure 1). These include: 2.1. water hyacinth occurrence data preparation, 2.2. covariate selection and preparation, 2.3. model tuning and cross-validation and 2.4. model explainability. In addition, details of software used in the analysis can be found in section 2.5.

Diagram

Description automatically generated

**Figure 1**: The general workflow used in this study to determine the relative importance of likely drivers of water hyacinth occurrence. Inputs are shown in green; outputs are shown in grey.

* 1. **Water hyacinth occurrence data preparation**

Based on previous work [9], the presence of water hyacinth across South Africa can be predicted with an accuracy of 80-93 %. The species distribution was derived using a Random Forest model calibrated on a combination of in-situ reference data and satellite-derived topographic, climate, weather, and Landsat-8 spectral variables [9]. The distribution was mapped at a spatial resolution of 30 m and was found to cover 2.69 % of the total (permanent and seasonal) surface water area within the country during 2013. This distribution corresponds to 27 206 infested waterbodies as detected, and provided within the GSW dataset for 2013 [30]. These infested waterbodies were subsequently labelled as the positive class (1, n = 27 206). The remaining waterbodies were all labelled as the negative class (0, n = 221 164).

* 1. **Covariate selection and preparation**

121 covariates were compiled for this study based on their availability as EO data products for South Africa and known influence on water hyacinth. Table 1 provides a list of the EO derived products used as covariates, their temporal coverage, and sources. The 121 potential drivers can be grouped as follows (Note: for reported links between driver and water hyacinth occurrence, refer to citation(s) below):

* Climate: Temperature [43–45], precipitation, frost [43], wind [12], and topo-climate [45,46].
* Social: Human modification and wealth index [47,48].
* Environment: Interspecies competition [49,50], distance to coastline [2,8,51], runoff, flood risk and surrounding nutrients [2,8].
* Land Use/Cover: Broad and fine class surrounding landcover.
* Topography: Elevation [25], including numerous morphology indices and landform classes.
* Hydrography: River connectivity and water seasonality [12,50].

Some variables were not available for 2013, therefore the closest temporally available data was used instead. For those covariates that were not readily available within the GEE data catalog or GEE community datasets, the datasets were downloaded (refer to Table 1 for data sources) for South Africa and uploaded to GEE.

Considering the large number of available environmental layers (n = 140), a standardized and reproducible procedure to select covariates to be used for modelling was implemented to 1) reduce redundancy between covariates, 2) obtain a more parsimonious and computationally efficient model, 3) to decrease the risk of over-fitting, 4) to avoid a biased assessment of variable importance, 5) to prevent the curse of dimensionality from negatively affecting the model performance and 6) to make the model easier to interpret [52,53]. The covariate selection procedure consisted of three steps: removal of irrelevant features, removal of redundant features through de-correlation and the removal of features with a low predictive power through Recursive Feature Elimination.

**2.2.1. Step 1: Removal of irrelevant features**

The number of consecutive nights that experienced temperatures less than 10°C and the number of upstream rivers was removed owing to low variance across waterbodies. Other irrelevant features such as snow and moss cover fraction from the global broad class landcover data were removed since they are not applicable to South Africa.

**2.2.2. Step 2: De-correlation analysis**

De-correlation analysis was carried out to reduce the redundancy of information between the remaining 136 environmental layers. Only covariate layers that had a low pairwise correlation coefficient <= 0.7 [54] with all other covariates were included in the subsequent analyses. For each pair of covariates correlated above this threshold, a single covariate was manually selected. Variables with higher spatio-temporal resolution and availability were prioritised and passed (n = 103) to the final stage of feature selection.

**2.2.3. Step 3: Recursive Feature Elimination**

Recursive feature elimination and cross-validated selection (RFECV) is a method that has proven effective to select an optimal set of covariates for tree-based and linear models (for example, [55,56]). The algorithm starts by fitting a model using all covariates, assessing model performance, and ranking covariate importance. The least important covariates are then removed from the pool, and again the model is fitted, assessed, and the least important covariates removed. The procedure repeats down to a single covariate. RFECV resulted in 82 final features for modelling as indicated in Table 1.

* 1. **Cross-validation, model selection and model tuning.**

Owing to the high spatial variation in (un)infested observation density and inherent spatial autocorrelation, randomly splitting the observations into k validation folds may result in a large difference between infested and infested instances (i.e., severe class imbalance) and high levels of spatial autocorrelation between folds, and therefore inflated model skill [57–59]. Therefore, a block cross-validation (CV) strategy was adopted whereby the presence/absence observations are spatially grouped by 5 km blocks. The blocks were randomly assigned to one of ten folds. In this way, all observations (presences or absences) belonging to a block were always in the same fold (not separated) during model calibration and validation. This ensures a more balanced positive-negative class distribution and a lower level of spatial autocorrelation in each fold. Thereafter, training occurred on 9 folds while validation was carried out on the remaining fold. This was performed until each of 10 folds were used as a validation fold (i.e., block CV). Model selection, feature selection, hyperparameter tuning, model evaluation and model interpretation was performed using a block CV strategy.

Random forest is an ensemble algorithm that endeavours to make an accurate classifier from multiple weak classifiers (decision trees) [60]. This is done by randomly selecting a subset of explanatory variables and using each subset to fit different decision tree models. Thereafter, combining the results using a majority vote. In this study, random forest was selected after evaluating 15 candidate models (Table 2). Precision was selected as the criteria for model selection since the cost of false positives are high i.e., this study aims to understand the occurrence of water hyacinth. The selected random forest model had the highest mean validation-fold precision score (Table 2). The optimal model hyperparameters were evaluated using a sequential model-based optimisation strategy.

Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC) and balanced accuracy are preferred metrics for the evaluation of models fitted on imbalanced datasets since these metrics are less sensitive to differences in the number of positive and negative cases i.e., class imbalance [61]. All metrics, except MCC (-1 to 1) range from zero to one, where a higher value is indicative of a more accurate model. Precision represents the ratio of true positives to total positives (true and false positives). Recall represents the ratio between true positives and the total actual positives (true positives and false negatives). Precision is maximised when the cost of false positives is higher than the cost of false negatives. In contrast, recall is maximised when the cost of false negatives is associated with higher costs than false positives. However, if both false positives and negatives have similar costs, F1 score should be maximised. Lastly, MCC incorporates both true and false positives and true and false negatives.

**Table 1:** Variable descriptions, associated units, temporal coverage and data source of the explanatory variables considered to investigate the likely drivers of water hyacinth occurrence. All explanatory variables were downloaded at a 30m spatial resolution. Those variables that were available at a coarser spatial resolution (> 30 m) were automatically resampled to 30 m using bilinear interpolation within Google Earth Engine.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Units | Temporal coverage | Source |
| Number of consecutive nights with less than 10 degrees Celsius | days | 2013 | MODIS LST- night |
| Area of waterbody | m2 | 2013 | [9] |
| Median river width at mean discharge | m |  | [62] |
| Area of aquatic vegetation | m2 | 2013 | [9] |
| \*Minimum Temperature (degrees Celsius) | °C | 1970 - 2000 | WorldClim |
| Connectivity- upstream and downstream river count | river count | 2000 | WWF hydrosheds [32] |
| \*Water seasonality- number of months water is present | month count | 1984 - 2019 | JRC GSW [30] |
| Total precipitation | mm | 2013 | TerraClimate |
| \*Distance to nearest coastline | m | 2009 | NOAA (https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/) |
| Distance to roads | m |  |  |
| Elevation | m | 2000 | NASADEM |
| \*Continuous Heat Insolation Load Index (0=cool, 255-warm). Surrogate for effects of shading and topographic insolation | 0-255 | 2006 - 2011 | SRTM, [63] |
| \*Global Human Modification (1= high modification) | 0-1 | 2016 | [31] |
| \*Topographic diversity- Surrogate for the variety of temperature and moisture conditions available to species as local habitats | -1323 - 8.81 | 2006 - 2011 | SRTM, [63] |
| Frost duration - Median duration of frost | day count | 2007 | [64] |
| Number of days below 0°C (0-365) | day count | 2013 | MODIS- LST, night temperature |
| Number of days below 10°C (0-365) | day count | 2013 | MODIS- LST, night temperature |
| \*South African National Land Cover (73 class) | km2 | 2018 | GeoTerraimage |
| \*Mean Annual Runoff | mm/year | 2005 | Strategic Water Source Areas (SWSA) |
| Broad (10) class Landcover | % (coverFraction) | 2015 | Corine Landcover |
| \*Flood hazard with a 10-year return period | water depth (m) | 2016 | Joint Research Commission [65] |
| Riparian soil nitrogen (5km buffer) | g/kg | 1905 - 2016 | Soil grids |
| Riparian soil ph (5km buffer) | pH | 1905 - 2016 | Soil grids |
| \*Riparian soil carbon (5km buffer) | g/kg | 1905 - 2016 | Soil grids |
| Stream Power Index (SPI) | 0-1 |  |  |
| Relative wealth index | -1 - 1 (high) |  | Facebook |
| 13 iSDA Soil layers (includes total nitrogen, etc.) | various |  | [66] |
| Mean wind speed for a 10 year period | m/s | 2008-2017 | Global wind atlas |
| \* Included in final model | | | |

**Table 2:** Evaluation metrics for 15 candidate models calculated based on a 10-fold block cross-validation strategy sorted by precision. The top-most entry highlights the performance of the random forest model after feature selection but prior to hyperparameter tuning. The highest scoring metric (refer to Section 2.3 for metric descriptions) is highlighted in yellow.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Recall** | **Prec.** | **F1** | **MCC** | **Balanced accuracy** |
| Random Forest Classifier | 0,5823 | 0,8269 | 0,6831 | 0,6621 | 0,7832 |
| Extra Trees Classifier | 0,5827 | 0,8182 | 0,6801 | 0,6580 | 0,7829 |
| CatBoost Classifier | 0,6252 | 0,7906 | 0,6981 | 0,6696 | 0,8018 |
| Gradient Boosting Classifier | 0,4338 | 0,7737 | 0,5555 | 0,5417 | 0,7086 |
| Extreme Gradient Boosting | 0,6100 | 0,7725 | 0,6815 | 0,6511 | 0,7933 |
| Light Gradient Boosting Machine | 0,5762 | 0,7660 | 0,6573 | 0,6276 | 0,7766 |
| Ridge Classifier | 0,1760 | 0,6941 | 0,2798 | 0,3134 | 0,5828 |
| Ada Boost Classifier | 0,4433 | 0,6757 | 0,5351 | 0,5017 | 0,7077 |
| Logistic Regression | 0,3094 | 0,6272 | 0,4136 | 0,3930 | 0,6426 |
| Decision Tree Classifier | 0,6127 | 0,6009 | 0,6065 | 0,5547 | 0,7798 |
| Linear Discriminant Analysis | 0,3869 | 0,5979 | 0,4691 | 0,4286 | 0,6764 |
| SVM - Linear Kernel | 0,4365 | 0,4695 | 0,3901 | 0,3464 | 0,6662 |
| K Neighbours Classifier | 0,2698 | 0,4636 | 0,3409 | 0,2923 | 0,6145 |
| Quadratic Discriminant Analysis | 0,5887 | 0,4151 | 0,4865 | 0,4148 | 0,7402 |
| Naive Bayes | 0,5244 | 0,3852 | 0,4438 | 0,3643 | 0,7075 |

* 1. **Model explainability**

To determine the contribution of each variable to the occurrence of water hyacinth, SHAP is applied to the entire dataset using the block CV strategy. This provides a contribution value for each predictor value in the dataset. This allows us to:

1. aggregate SHAP values by predictor and therefore rank the overall importance of variables for South Africa (Figure 3) or for sub-groups (for example, provinces, Figure 4) using summary plots,

2. understand the effect of individual variables on the probability of water hyacinth occurrence using SHAP (main effect) dependence plots, or

3. to understand variable interaction effects on the probability of water hyacinth occurrence using standard SHAP dependence plots (Figure 5), and

4. rank SHAP values for an instance (waterbody) to determine the contribution of each variable on the occurrence of water hyacinth at the site using waterfall and contribution plots.

Each point in a SHAP summary plot represents a SHAP contribution value towards the probability of water hyacinth occurrence at a site, while an aggregate of points indicates the magnitude, commonality, and direction of the variable's global effect. The mean SHAP value across all sites represents a measure of global variable importance (Figure 3). Predictors are listed in the order of global importance, from top to bottom**.**

A SHAP dependence plot consists of points that represent a variable's SHAP value for a given site. Dependence plots provide information on the variance of variable importance in response to changes in the variable's value. Vertical dispersions in SHAP dependence plots (Figure 5, left) depict the magnitude of interaction effects. The points are coloured by the feature which explains the most variation among the remaining modelled features.

One of the benefits of SHAP dependence plots over traditional partial dependence plots is the ability to distinguish between models with and without interaction terms. In other words, SHAP dependence plots give an idea of the magnitude of the interaction terms through the vertical variance of the scatter plot at a given feature value. SHAP dependence plots may also provide insight into the effect of a feature on the outcome without considering the interaction effects of any other feature.

* 1. **Implementation details**

All analysis were achieved using python-based packages and/or the GEE python Application Programming Interface (API). The environmental covariate data was automatically extracted (in batches) for all GSW detected waterbodies within South Africa using the Google Earth Engine python API. This was facilitated by the geemap package [67]. RFECV feature selection was accomplished using the scikit-learn package. Model selection was achieved using the pycaret library [68]. Hyperparameter tuning was achieved using the hyperopt package [69]. Model explainability was achieved using the shap package [39]. All figures and maps were created using matplotlib [70], seaborn [71], geopandas [33] and contextily [72] for basemaps.

1. **Results**

The top three factors influencing the occurrence of water hyacinth is minimum temperature in the coldest month, distance from the coast and water seasonality. The order of variable importance varies by province. We show the spatial distribution of variable importance across South Africa at a 5 km spatial scale. Lastly, the contribution of variables for individual waterbodies have been quantified. These findings are based on the interpretability of a good performing model (F1 > 0.7) using current best spatial modelling practice that promotes interpretation consistency.

* 1. **Model evaluation**

**Chart

Description automatically generated**Model accuracy has implications on the reliability and consistency of model explanations. The model shows overall good performance (F1 score > 0.7, as defined by [38]). However, the model does not perform as well based on Matthews Correlation Coefficient (mean = 0.49) since the model finds it difficult to detect sites that do not have water hyacinth based on the suite of explanatory variables considered in this study. Moreover, the narrow and long (blue area) distribution highlights the high variability in MCC in contrast to the other metrics. The model precision is the highest scored metric (mean = 0.814), corresponding to 81.4% of water hyacinth infested sites being correctly identified. This is advantageous when trying to understand the occurrence (as opposed to the absence) of a species.

**Figure 2:** The distribution of the final random forest model evaluation metrics based on 10-fold block cross-validation. Highlighting a good overall performance based on the F1-score and a poorer performance based on the Matthews Correlation Coefficient.

* 1. Local (per waterbody) interpretation

To determine the contribution of each of the 17 most important variables to the occurrence of water hyacinth at a site, a waterfall plot can be used (Figure 3). In this case, the water hyacinth infested Roodeplaat dam has a SHAP value of 0.437, a value much greater than the expected base value (-4.299) i.e., the average SHAP value across all sites in South Africa. Since this log odds value is greater than 0, the probability is greater than 0.5 that water hyacinth could occur at this site. Minimum temperature (2.9 °C) in the coldest month between 1970- 1990 and the persistence of surface water (surface water present for more than 9 months of 2013) are the two main factors promoting (dark pink) the occurrence of water hyacinth at Roodeplaat dam.

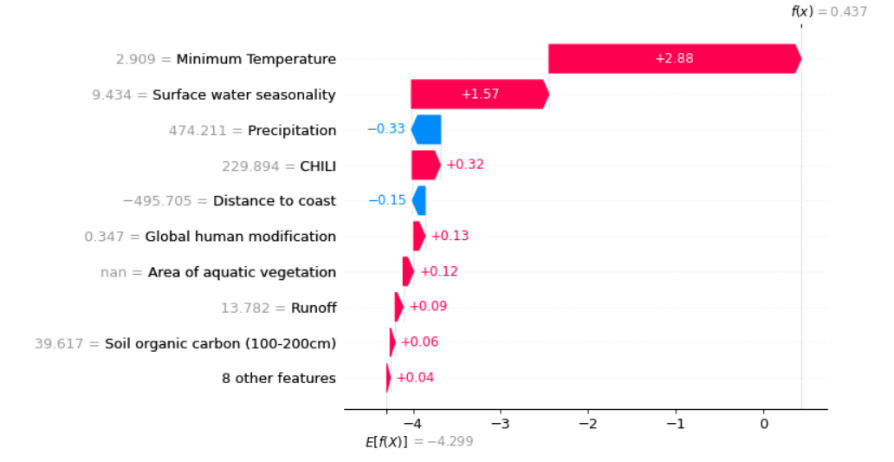
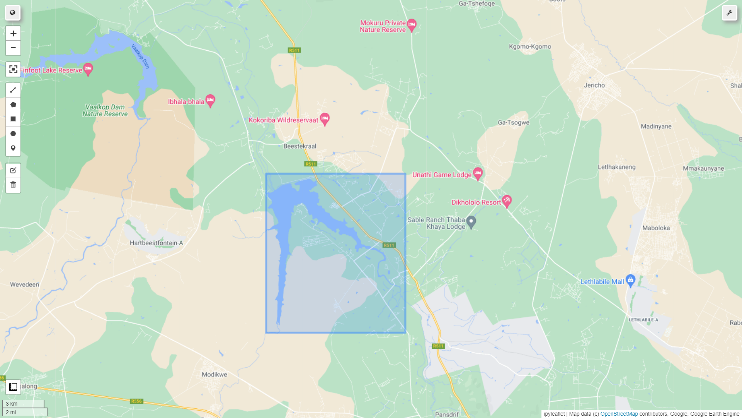


Figure 3: The location of Roodeplaat dam, South Africa and the associated variables’ contributions to the occurrence of water hyacinth at this dam.

* 1. **Global feature importance**

The best (top-most) predictor of water hyacinth presence is the minimum temperature as based on the bioclimatic variable’s minimum temperature in the coldest month (Figure 4). This is indicated by the clusters of pink and blue dots representative of sites with warmer minimum temperatures and colder minimum temperatures respectively, formed at either ends of the probability of water hyacinth distribution. In addition to the commonly acknowledged critical role of low temperatures on limiting water hyacinth distribution [43,73], insights into less researched environmental gradient effects on water hyacinth are highlighted (Figure 4 and 5). For example, the occurrence of artificial surfaces (industrial and commercial), bare ground (fallow lands and old-fields, low-intermediate grass cover), and viticulture fields, often irrigated (cultivated commercial permanent vines) are all associated with a higher probability of water hyacinth. Water systems associated with low shading effects, and a heterogenous temperature and moisture i.e., dry to humid topo-climate niche indicated by high CHILI and Topographic diversity values also improve the suitability for water hyacinth occurrence.

The differences in ranking of overall variable importance between provinces reinforces the need for considering the social and ecological contexts that likely influences the probability of successful management (Figure 5). Within the Western Cape, runoff is the fourth most important factor determining the occurrence of water hyacinth while in the Highveld region of Gauteng, it ranks much lower (ninth position). Similarly, grass cover fraction has been identified as an important variable in Gauteng (fourth) while, in the Western Cape grass cover fraction has been ranked at tenth position.

A picture containing chart

Description automatically generated**Figure 4:** ASHAP summary plot for the top 17 features and their direction of effect, in order of importance, from top-to-bottom, for the prediction of water hyacinth presence across South Africa.

**A picture containing text

Description automatically generated**A picture containing diagram

Description automatically generated

**Figure 5:** SHAP summary plots for the top 17 features and their direction of effect for the Western Cape province (left) and the Gauteng province (right), in order of importance, from top-to-bottom, for the prediction of water hyacinth presence. Gaps are attributed to the unavailability of environmental conditions

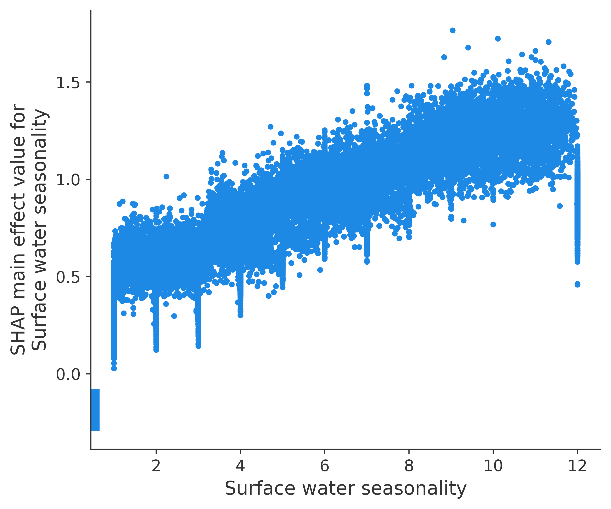
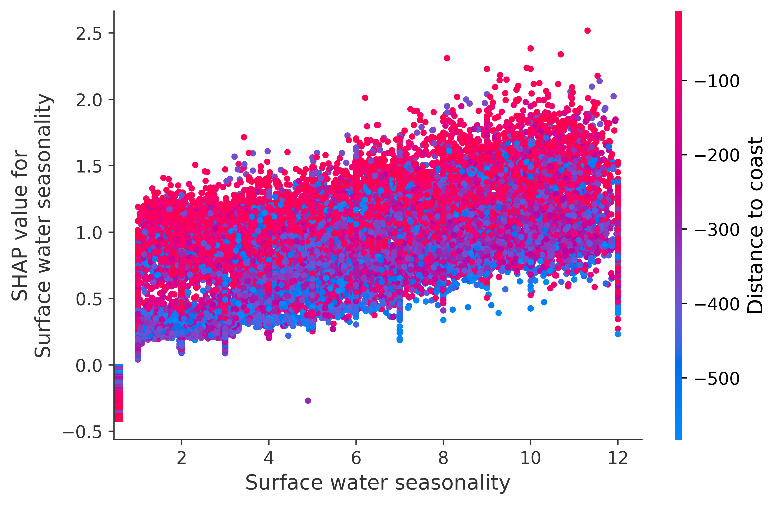
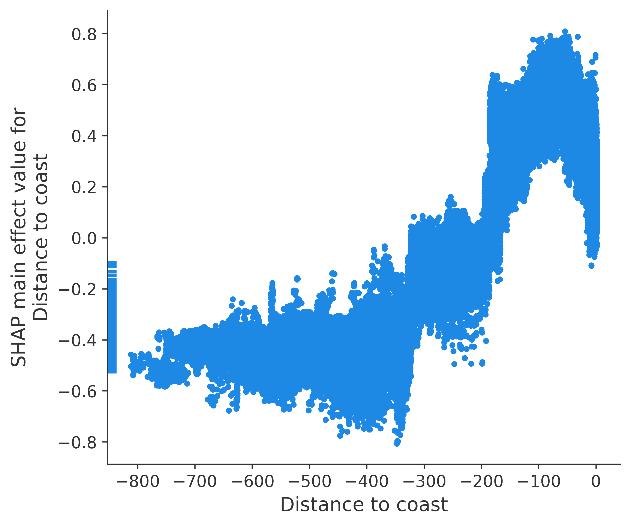
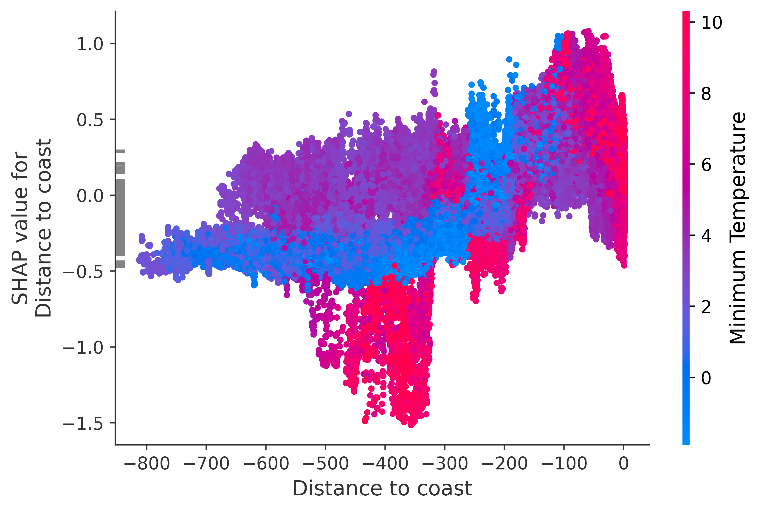
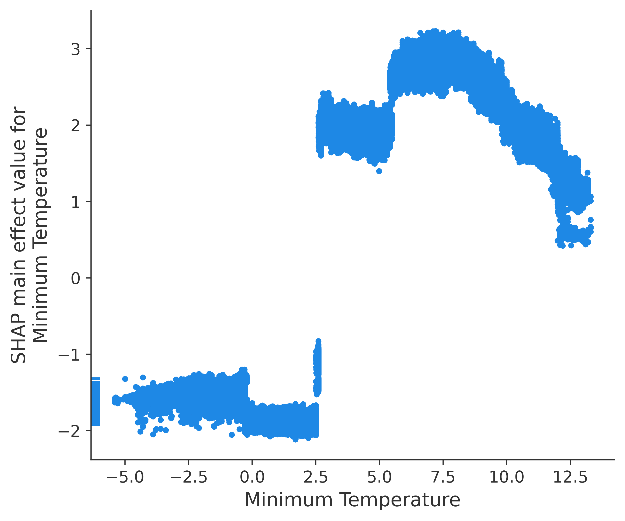
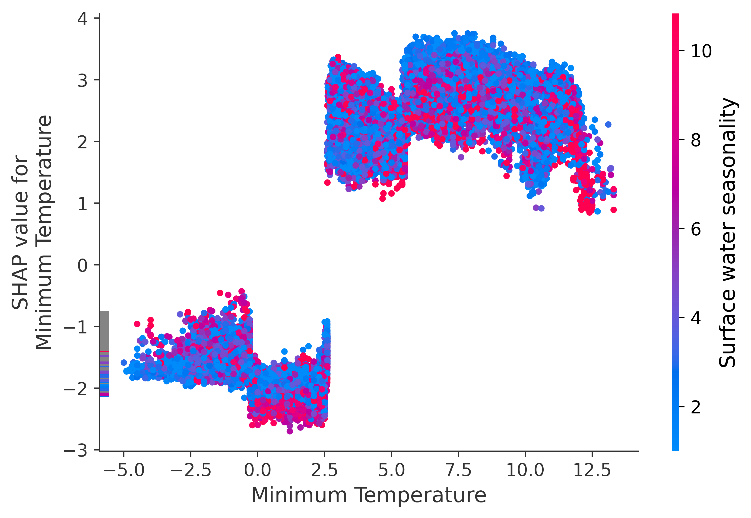
* 1. **Variable dependence and interactions**

Many important pairwise interactions included minimum temperature and distance from the coast. These were also the predictors with the highest global feature importance (Figure 4). By separating the vertical dispersions in the dependence plots that represent interaction effects (pre-computed by SHAP), the sole contribution of a variable to the predictions was obtained, as shown in the main effect plots (Figure 6b, 6d, 6f).

There are abrupt changes in variable importance at ~ 2.5 °C and 5 °C that both favour the occurrence of water hyacinth with the former inflection being more important for water hyacinth i.e., the SHAP values exceed 0 at temperatures greater than 2.5 °C. Surface water persistence has a gradual positive trend with the occurrence of water hyacinth. Longer surface water persistence (values along the x axis) combined with shorter distances to the coast (red colours) increase the suitability of a water system for water hyacinth (Figure 6e and 6f).

**With Interactions**

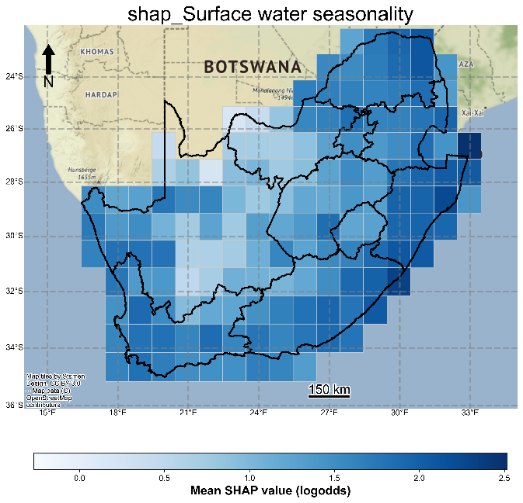
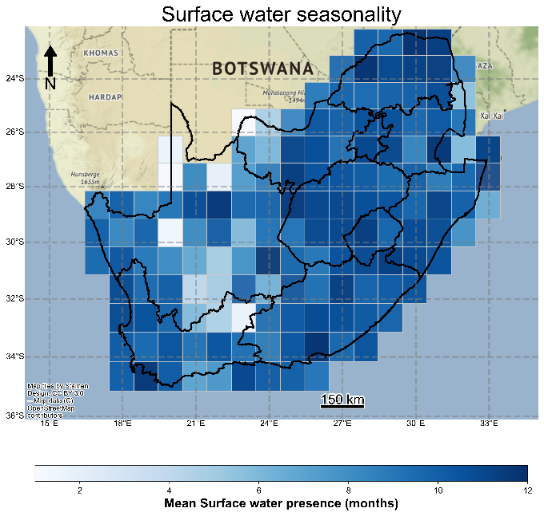
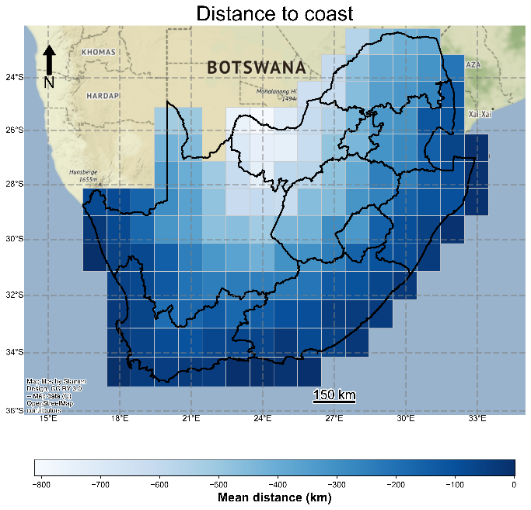
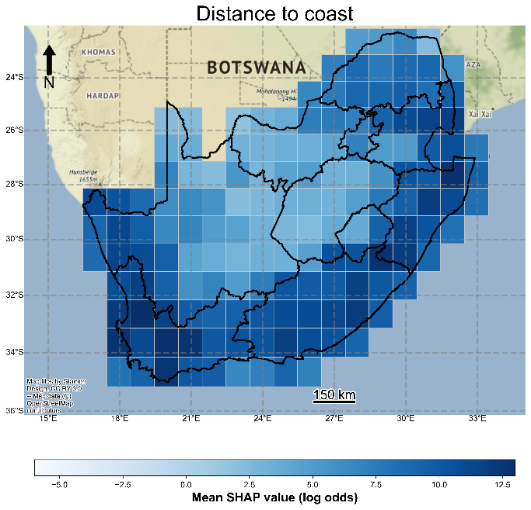
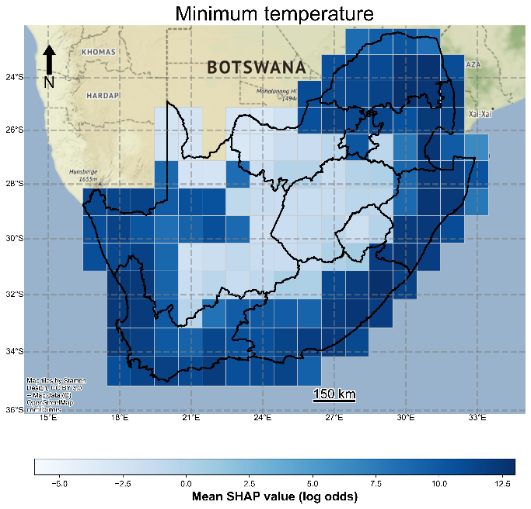
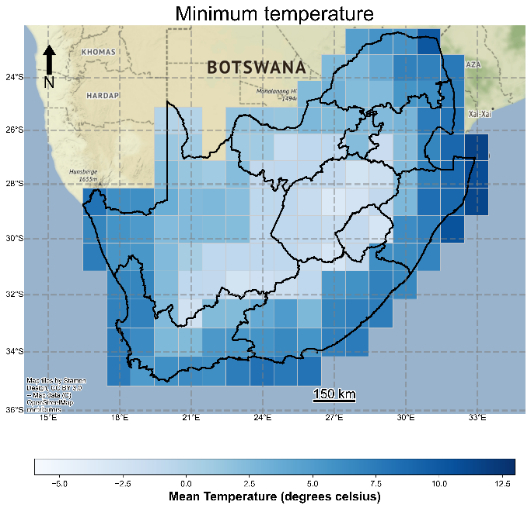
**Main Interactions**



**Figure 6:** SHAP dependence plots for the top three predictors of water hyacinth with variable interactions (a, c, e) and main interactions (b, d, f). Minimum temperature refers to the coldest temperature in the coldest month between 1970-1990. A value of zero kilometers from the coast refers to pixels at the coast, while a value of -800 refers to a pixel 800 km from the coast. Water persistence indicates the average number of months the pixels of a waterbody contain water.

* 1. **Spatial distribution of variable importance**

To capture the spatial distribution of a variable’s importance, the variables’ SHAP values were aggregated at a ~5 km block level (Figure 7) by taking the mean SHAP value. Low temperatures are associated with a low probability of water hyacinth occurrence. Moreover, the odds of water hyacinth occurrence has a much more abrupt change (Figure 7b) in relation to the temperature gradient (Figure 7a). Both the distance from the coast and surface water persistence have a less abrupt effect on the probability of water hyacinth occurrence. The unsuitability of the interior of the country can also be noted. Refer to the supplementary material for the spatial distribution of the remaining variables’ importance.



**Figure 7:** The mean SHAP value distribution of the top three predictive features across South Africa.

Within the interior of the country, there is greater variability in the factors that determine the occurrence of water hyacinth in comparison to the coastal regions (Figure 8). Minimum temperature is most important within the coastal regions, while runoff is more important for areas in the interior region of the country. For the areas outside coastal regions in the south, the distance from the coast is important.

Map

Description automatically generated

Figure 8: The most important predictor of water hyacinth occurrence per block (~111 km) across South Africa.

1. **Discussion**

The most appropriate management strategy is not always obvious and is dependent on how an IAAP responds under different environmental and socio-economic contexts [12]. This study used pre-existing EO-derived datasets, SDMs and xAI to infer the relative importance of range and niche defining variables of water hyacinth within South Africa at multiple spatial scales. In addition to reporting novel quantitative insights based on a national extent analysis of water hyacinth occurrence, many of the species-environment interactions of water hyacinth align with known thermal limits of the plant, and environmental plant preferences. Beyond revealing general correlations between water hyacinth occurrence and environmental and socio-economic variables, we also analysed these correlations separately for different provinces, and sites. With this information, site-specific management strategies may be prioritised in a cost-effective and data-driven approach.

* 1. **The influence of climatic factors on water hyacinth distribution**

The interaction between temperature and water hyacinth occurrence has been the most widely investigated variable among those considered in this study. Given the importance of this variable against all the other variables considered in this study, the interest in understanding minimum temperature effects on water hyacinth is warranted. For the purposes of this study, the previously identified range-limiting physiological temperatures for water hyacinth growth can be considered to verify the models’ derived water hyacinth-temperature interactions.

* + 1. **Temperature and frost**

Below 2.5°C, overwintering of water hyacinth is extremely unlikely. Air temperatures < 0°C significantly increase the mortality of the plant by killing above-water portions [74]. This is supported by the negative SHAP values and the sharp decrease in the log odds of water hyacinth presence below a minimum temperature of 2.5°C Figure 6, variable 1). Such areas likely experience repeated exposures to severe winters that limit the plant’s range [45]. This may result in smaller populations of overwintering water hyacinth surviving to the following season, such that nuisance populations may not form in areas along the limiting cold (boundary) edge of the IAAPs distribution. The cold edge is likely to lie at the boundary between contrasting blocks of low and high SHAP values (Figure 7b).

Between 2.5 and 5°C, water hyacinth may be able to persist (indicated by the positive SHAP values, Figure 6, variable 1). However, these will be smaller populations if the low temperature conditions are short-lived. Water hyacinth is mainly limited by its winter-survival temperature at around 7 °C, although a short period of exposure at 5 °C may not entirely destroy the plant [44]. In addition, the persistence of water hyacinth at these temperatures may be attributed to the ineffectiveness of biological control agents. Low temperature conditions and the lower plant quality may prevent agent establishment or suppress insect population growth directly by slowing down their development [43,75,76]. Water hyacinth can tolerate air temperatures less than 5°C for a limited period but will experience a steady decline in regrowth potential [44]. This is supported by the dependence plot for minimum temperature (Figure 6a and 6b), a dramatic vertical increase at 5 degrees Celsius can be observed. However, there is a gradual decline in the log odds of water hyacinth presence at minimum temperatures greater than 5-8 degrees Celsius. This may be attributed to the proliferation of water hyacinth, its subsequent management prioritisation, and the increased effectiveness of biocontrol agents at warmer temperatures. This suggests the need to prioritise biological control at sites with a minimum temperature above 8°C i.e., beyond the peak of water hyacinth suitability. Mechanical and chemical control management strategies should be prioritised for areas with minimum temperatures lower than 8 °C.

For example, cold-stress limits the range of water hyacinth in a few small high elevation areas in South Africa [25]. Water hyacinth is reported to be sensitive to frost. Frosts kill the leaves and upper petioles that protect the rhizome, but prolonged cold temperatures, below 5°C, may also kill the rhizome, resulting in death of the plants [65]. The ability of the seeds and bulbs to survive winter conditions is likely to be a range-limiting factor.

* + 1. **Microclimate effects**

Those areas that are concurrently further from the coast (> 400 km) and have a high Continuous Heat Insolation Load Index (CHILI >200) value (Figure S3, variable 9), associated with warmer topo-climates attributed to the absence of topographic shading, exhibit a higher suitability for water hyacinth. These model derived insights coincide with the preference of water hyacinth for warmer regions. In comparison to CHILI, topographic diversity shows a less variable effect on water hyacinth occurrence and shows a gradual positive trend with water hyacinth occurrence.

When considering the combined effects with high rates of runoff (>75 mm/year) there is a divergent effect on water hyacinth habitat suitability. Regions with low topographic diversity and high rates of runoff are associated with habitat unsuitability. In contrast, if both topographic diversity and runoff is high, the region promotes the occurrence of water hyacinth. Higher topographic diversity and therefore available topo-climate niches are associated with higher species resilience under climate change [46].

* + 1. **Precipitation**

New year’s dam is a small (150 ha), shallow, oligotrophic dam within the Eastern Cape. Above-average rainfall in this semi-arid area has been suggested to have initiated the weed’s resurgence in 1998 by increasing the nutrient input into the dam [76]. Based on the dependence plot for precipitation (Figure S3, variable 4), an accumulated precipitation greater than 800 mm is associated with an increased chance of water hyacinth. The effect of precipitation below 800 mm varies depending on the context of other variables.

Precipitation links to the presence and persistence of surface water. The dependence plot for water persistence shows a positive trend with the occurrence of water hyacinth (Figure S3). While water hyacinth is resilient to water level changes [77], waterbodies with high seasonality are less likely to support water hyacinth. Permanent waterbodies have four times the (log)odds of supporting water hyacinth compared to seasonal surface water that is only present for 1-3 months of the year. In the dry karoo areas of the Northern Cape, it is likely that there are few suitable sources of standing water to support water hyacinth. Simultaneously, the closer a waterbody is to the coastline, the greater the probability of water hyacinth occurrence despite water seasonality.

* 1. **Influence of socio-economic factors on water hyacinth distribution**

The presence of alien and invasive species is strongly linked to human activity through human aided dispersal for introduction and spread [48] and, human-induced disturbance for establishment [47]. Based on the human modification variable that captures human influence from 13 datasets representing human settlement, agriculture, transportation, mining, and energy production [31], there is a parabolic relationship with the presence of water hyacinth (Figure S3, variable 5), whereby the odds of water hyacinth occurring peaks at a human modification score of 0.35. Human modification values on either extreme limit the distribution of water hyacinth occurrence. Higher human modification may cause more disturbance than water hyacinth is able to tolerate, or alternatively, the plant may persist and because of its associated negative consequences, the plant may be targeted for removal.

* 1. **Influence of environmental factors on water hyacinth distribution**

The occurrence of other aquatic species largely reduces the probability of water hyacinth occurrence likely through interspecies competition and the reduction of available nutrients for water hyacinth proliferation (Figure S3, variable 11) [78,79]. Here, smaller populations (740-750 m2) of other aquatic plant species are associated with lower water hyacinth habitat suitability. This may be attributed to less suitable abiotic conditions that hinder water hyacinth out competing other species of IAAP. As a result, this may suggest that the co-occurrence of water hyacinth and other IAAP species may both be effects of underlying drivers as opposed to interspecies competition driving the low suitability for water hyacinth occurrence. For example, being closer to the coast increases the habitat suitability for water hyacinth, at which stage the occurrence of other aquatic vegetation species has little to no effect on water hyacinth occurrence (refer to Figure S3). Owing to the plant’s free-floating characteristics and not being limited to shallow water depths, water hyacinth is able to outcompete littoral vegetation [2]. The IAAPs’ phenotypic plasticity has also been reported to aid water hyacinths’ competitiveness. Established mats tend to extend their leaf area and increase in height by lengthening their petioles, likely shading out other IAAPs [49,80]. This coupled with the high vegetative reproduction on the leading edge of newly formed mats are probably the mechanisms that have aided in the documented ability of water hyacinth to out compete other floating weeds [49,78].

Since satellite-based estimations of water nutrient levels are not readily available, the soil nutrients for the (5 km) area surrounding a waterbody were used instead. However, all soil nitrogen, ph and Soil Organic Carbon (SOC) variables, except for the SOC (100-200 cm) were removed owing to high correlation (>0.7) with distance to the coast, runoff and precipitation. The SOC variable that remains shows a varied influence on water hyacinth occurrence. However, it is dominated by reduced odds of water hyacinth occurrence, especially for low levels of SOC.

Floods are associated with an increased chance of water hyacinth occurrence [81] (Figure S3, variable 12). Floods increase the dispersal ability of water hyacinth downstream and may also decrease the effectiveness of biological control at these sites [82], whereby biocontrol agent populations take much longer to recover than their host plant. Water hyacinth seeds buried in the substrate germinate under the now open water with access to sunlight. that may undergo rapid germination under conducive conditions facilitated by the now open water, access to sunlight and the inflow of nutrients [81,83]. Moreover, these results highlight the inability of floods to act as a regulatory mechanism for the 2013 water hyacinth population. Floods have been suggested as a mechanism to force plants into intolerable saline water conditions [2]. This is despite the high chances of occurring i.e., many water hyacinth infestations occur near coastal areas.

* 1. **The benefits and drawbacks**

Correlative SDMs remain valuable for IAAP species risk mapping and management but are often criticised for a lack of biological underpinning and incorrect modelling approaches. Modelling experts partly address the former concern by using prior expert knowledge of species’ requirements or tolerances to define likely and potential variables for modelling to reduce the evaluation of spurious species-environment interactions. Nevertheless, SDMs are correlative in nature and therefore, all insights drawn from this study need to be considered with caution, especially if species-environment feedback mechanisms occur. For the latter, some of the incorrect practices include not addressing correlation between variables, spatial autocorrelation, ignoring species-specific model selection, incorrect background data sampling or pseudo-absence data selection.

Highly correlated variables must be removed to allow accurate model interpretation. Spatial autocorrelation must be taken into consideration to avoid overestimation of model performance and avoid artefacts in output habitat suitability maps. This has been achieved by using block cross-validation during feature selection, model selection, model tuning, model evaluation and model interpretability. Since no single algorithm is the best for all SDMs, model selection should be carried out. Packages such as pycaret make this easy to implement and therefore easier to be adopted by the wider SDM community.

By using actual distribution data, the benefits of using presence-absence data over presence-background data used with Maxent or Ecological Niche Factor Analyses (ENFAs) or presence-pseudo-absence data used with other ML algorithms are inherited. This includes reduced uncertainty compared to using pseudo-absence data. High quality EO-derived distribution maps are more comprehensive and less prone to sample bias compared to costly field collected samples. These biases violate the assumption of independence among species records [84]. Consequently, the output habitat suitability maps based on biased data may not only correspond to the species’ observed distribution, but to the distribution of sampling effort. However, at the same time all EO-derived distributions are prone to mapping errors that may include residual spatial autocorrelation. Residual spatial autocorrelation is spatial autocorrelation in the distribution of model error, causing the model to be spatially-differentially reliable. As a result, these errors and limitations may be propagated to the interpretability of SDMs.

Mechanistic models require costly and labour-intensive field or lab experiments to obtain sufficient data for a representative set of drivers that allow for creating SDMs, and may still not hold across large national extents with varying climate regimes (for example, [12]). While many EO-derived datasets provide numerous readily available explanatory variables, the availability of these variables are constrained by technical limitations, spatio-temporal extensibility, poor accuracy, and the overall early stages of the EO sector. Moreover, it is not always possible to directly compare EO-derived values with instantaneous values derived through direct observations. For example, the minimum temperature data is dependent on long-term climate data and may not always correspond to instantaneous direct measurements. Nevertheless, the results in this study for minimum temperature seem to fit sufficiently well with known critical temperatures governing water hyacinth growth.

The overall performance of the SDM used in this study results in less than 30% error based on the F1-score. This may partly be attributed to the environmental conditions being suitable for water hyacinth occurrence, but water hyacinth not yet being introduced at these sites. If this is the case, it would suggest that the distribution of water hyacinth is strongly controlled by dispersal constraints, and thus might not be in complete equilibrium with its recipient environment throughout its entire South African range (see also, [85]). The error may also be an outcome of not including all variables (such as turbidity, water nutrient levels, history of management interventions, and water depth) that influence the occurrence of water hyacinth.

1. **Future work**

SHAP values represent the odds (after being log transformed) of, in this case, the occurrence of water hyacinth. Based on log odds alone it is difficult to translate these values into actionable insight for management. For this information to yield actionable insights more suited to guiding management interventions, counterfactual explanations may be useful via the DiCE python package [86]. Here, artificial data points with manually adjusted covariate values can be used to determine the effects on water hyacinth occurrence and therefore the effect of any proposed interventions. To reduce the probability of considering correlative species-environment interactions causal feature selection may be useful via the pyCausalFS package [87]. Lastly, considering the predictive power of non-spectral data used in this study, we recommend species mapping studies that rely on EO data to incorporate environmental data to improve the generalisation ability of ML models. Large scale EO-derived species distributions are rare; therefore, we investigated the tradeoff between data size and model accuracy. For a short discussion on dataset size refer to supplementary figure 2. The factors contributing to susceptibility, their spatial distribution and associated importance may aid in risk-prioritisation, justifying costly but more data rich monitoring exercises or traditional lab-controlled and artificial outdoor experiments. The knowledge gained from these experiments may, in turn, assist the development and improvement of model-based approaches. In this way, such an active learning-based approach may leverage the benefits from both traditional experimental setups and model-based methods that together promote the effective local management of introduced IAAP populations at an early stage, across the wide water hyacinth distribution.

1. **Conclusion**

The methods used in this study have demonstrated their ability to identify climate, microclimate, anthropogenic, topographic, and land use and land cover context favouring water hyacinth occurrence. Moreover, the species-environment insights highlight the value of combining EO, ML, xAI and GEE to understand the factors that limit and promote the establishment of water hyacinth and provides an indication of the areas where the IAAP is likely to extend its range in the future. This method may provide a data-driven and evidence-based option for the pre-selection and prioritisation of management strategies on a site-by-site basis. Owing to the negligible costs of carrying out this analysis, in comparison to similar large scale field studies, we hope to encourage similar computer-based studies and additional research into EO-derived species distributions that this study relies on.

1. **Data and code availability**

The dataset used in this study with water hyacinth occurrence and the extracted covariates has been published on Kaggle. The notebooks used in this study that allow for the complete reproducibility of this study, figures and tables are available from this GitHub repository. Moreover, notebooks with example analysis have been made available on Kaggle to assist researchers repeat the methods used in this study.

1. **References**

1. Julien, M.H. and Orapa, W. Successful biological control of water hyacinth (Eichhornia crassipes)in Papua New Guinea by the weevils Neochetina bruchi and Neochetina eichhorniae (Coleoptera: curculionidae). In Proceedings of the International Symposium on Biological Control of Weeds; Bozeman, MT, USA: Montana State University, 1999; p. 1027.

2. Coetzee, J.A.; Hill, M.P.; Ruiz-Téllez, T.; Starfinger, U.; Brunel, S. Monographs on invasive plants in Europe N° 2: Eichhornia crassipes (Mart.) Solms. *Bot. Lett.* **2017**, *164*, 303–326.

3. Pellegrini, M.O.O.; Horn, C.N.; Almeida, R.F. Total evidence phylogeny of Pontederiaceae (Commelinales) sheds light on the necessity of its recircumscription and synopsis of Pontederia L. *PhytoKeys* **2018**, 25.

4. Villamagna, A.M.; Murphy, B.R. Ecological and socio‐economic impacts of invasive water hyacinth (Eichhornia crassipes): a review. *Freshw. Biol.* **2010**, *55*, 282–298.

5. Vila, M.; Ibáñez, I. Plant invasions in the landscape. *Landsc. Ecol.* **2011**, *26*, 461–472.

6. Miller, B.E.; Coetzee, J.A.; Hill, M.P. Mind the gap: the delayed recovery of a population of the biological control agent Megamelus scutellaris Berg.(Hemiptera: Delphacidae) on water hyacinth after winter. *Bull. Entomol. Res.* **2021**, *111*, 120–128.

7. Tipping, P.W.; Smith, M.C.; Lake, E.C.; Minteer, C.R.; Goode, A.B.C.; Foley, J.R.; Gettys, L.A. Classical biological control and apparent competition: Evaluating a waterhyacinth invaded community module. *J. Appl. Ecol.* **2020**, *57*, 926–935.

8. Bick, E.; de Lange, E.S.; Kron, C.R.; da Silva Soler, L.; Liu, J.; Nguyen, H.D. Effects of salinity and nutrients on water hyacinth and its biological control agent, Neochetina bruchi. *Hydrobiologia* **2020**, *847*, 3213–3224.

9. Singh, G.; Reynolds, C.; Byrne, M.; Rosman, B. A Remote Sensing Method to Monitor Water, Aquatic Vegetation, and Invasive Water Hyacinth at National Extents. *Remote Sens.* **2020**, *12*, 4021.

10. Hill, M.P.; Coetzee, J. The biological control of aquatic weeds in South Africa: Current status and future challenges. *Bothalia-African Biodivers. Conserv.* **2017**, *47*, 1–12.

11. Mehta, S. V; Haight, R.G.; Homans, F.R.; Polasky, S.; Venette, R.C. Optimal detection and control strategies for invasive species management. *Ecol. Econ.* **2007**, *61*, 237–245.

12. Wilson, J.R.; Holst, N.; Rees, M. Determinants and patterns of population growth in water hyacinth. *Aquat. Bot.* **2005**, *81*, 51–67.

13. King, A.M. The effect of temperature on biological control of water hyacinth, Eichhornia crassipes (Pontederiaceae) in South Africa 2011.

14. Rainford, J.; Crowe, A.; Jones, G.; van den Berg, F. Early warning systems in biosecurity; translating risk into action in predictive systems for invasive alien species. *Emerg. Top. Life Sci.* **2020**, *4*, 453–462, doi:10.1042/ETLS20200056.

15. Kilroy, C.; Snelder, T.H.; Floerl, O.; Vieglais, C.C.; Dey, K.L. A rapid technique for assessing the suitability of areas for invasive species applied to New Zealand’s rivers. *Divers. Distrib.* **2008**, *14*, 262–272.

16. Joshi, C.; de Leeuw, J.; van Duren, I.C. Remote sensing and GIS applications for mapping and spatial modelling of invasive species. In Proceedings of the Proceedings of ISPRS; 2004; Vol. 35, p. B7.

17. Zhang, H.K.; Roy, D.P.; Yan, L.; Li, Z.; Huang, H.; Vermote, E.; Skakun, S.; Roger, J.-C. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. *Remote Sens. Environ.* **2018**, *215*, 482–494.

18. Huang, C.; Asner, G.P. Applications of remote sensing to alien invasive plant studies. *Sensors* **2009**, *9*, 4869–4889.

19. Müllerová, J.; Pergl, J.; Pyšek, P. Remote sensing as a tool for monitoring plant invasions: Testing the effects of data resolution and image classification approach on the detection of a model plant species Heracleum mantegazzianum (giant hogweed). *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *25*, 55–65.

20. Bradley, B.A. Remote detection of invasive plants: a review of spectral, textural and phenological approaches. *Biol. Invasions* **2014**, *16*, 1411–1425.

21. Elith, J. Predicting distributions of invasive species. *Invasive species Risk Assess. Manag.* **2017**, *10*.

22. Briscoe Runquist, R.D.; Lake, T.A.; Moeller, D.A. Improving predictions of range expansion for invasive species using joint species distribution models and surrogate co‐occurring species. *J. Biogeogr.* **2021**.

23. Barbet-Massin, M.; Rome, Q.; Villemant, C.; Courchamp, F. Can species distribution models really predict the expansion of invasive species? *PLoS One* **2018**, *13*, e0193085.

24. Kumschick, S.; Wilson, J.R.U.; Foxcroft, L.C. A framework to support alien species regulation: the Risk Analysis for Alien Taxa (RAAT). *NeoBiota* **2020**, *62*, 213.

25. Kriticos, D.J.; Brunel, S. Assessing and managing the current and future pest risk from water hyacinth,(Eichhornia crassipes), an invasive aquatic plant threatening the environment and water security. *PLoS One* **2016**, *11*, e0120054.

26. Rodriguez-Merino, A.; Garcia-Murillo, P.; Cirujano, S.; Fernandez-Zamudio, R. Predicting the risk of aquatic plant invasions in Europe: How climatic factors and anthropogenic activity influence potential species distributions. *J. Nat. Conserv.* **2018**, *45*, 58–71.

27. Kass, J.M.; Muscarella, R.; Galante, P.J.; Bohl, C.L.; Pinilla‐Buitrago, G.E.; Boria, R.A.; Soley‐Guardia, M.; Anderson, R.P. ENMeval 2.0: redesigned for customizable and reproducible modeling of species’ niches and distributions. *Methods Ecol. Evol.*

28. Ryo, M.; Angelov, B.; Mammola, S.; Kass, J.M.; Benito, B.M.; Hartig, F. Explainable artificial intelligence enhances the ecological interpretability of black‐box species distribution models. *Ecography (Cop.).* **2021**, *44*, 199–205.

29. Cha, Y.; Shin, J.; Go, B.; Lee, D.-S.; Kim, Y.; Kim, T.; Park, Y.-S. An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *J. Environ. Manage.* **2021**, *291*, 112719.

30. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422.

31. Kennedy, C.M.; Oakleaf, J.R.; Theobald, D.M.; Baruch‐Mordo, S.; Kiesecker, J. Managing the middle: A shift in conservation priorities based on the global human modification gradient. *Glob. Chang. Biol.* **2019**, *25*, 811–826.

32. Grill, G.; Lehner, B.; Thieme, M.; Geenen, B.; Tickner, D.; Antonelli, F.; Babu, S.; Borrelli, P.; Cheng, L.; Crochetiere, H. Mapping the world’s free-flowing rivers. *Nature* **2019**, *569*, 215–221.

33. Jordahl, K. GeoPandas: Python tools for geographic data. *URL https//github. com/geopandas/geopandas* **2014**.

34. Braunisch, V.; Coppes, J.; Arlettaz, R.; Suchant, R.; Schmid, H.; Bollmann, K. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography (Cop.).* **2013**, *36*, 971–983.

35. Bradie, J.; Leung, B. A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *J. Biogeogr.* **2017**, *44*, 1344–1361.

36. Effrosynidis, D.; Arampatzis, A. An evaluation of feature selection methods for environmental data. *Ecol. Inform.* **2021**, *61*, 101224.

37. Smith, A.B.; Santos, M.J. Testing the ability of species distribution models to infer variable importance. *Ecography (Cop.).* **2020**, *43*, 1801–1813.

38. Liu, B.; Udell, M. Impact of Accuracy on Model Interpretations. *arXiv Prepr. arXiv2011.09903* **2020**.

39. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.

40. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable machine learning for scientific insights and discoveries. *Ieee Access* **2020**, *8*, 42200–42216.

41. Wang, R.; Kim, J.-H.; Li, M.-H. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Sci. Total Environ.* **2021**, *761*, 144057.

42. Yu, H.; Cooper, A.R.; Infante, D.M. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecol. Modell.* **2020**, *432*, 109202.

43. Byrne, M.; Hill, M.; Robertson, M.; King, A.; Katembo, N.; Wilson, J.; Brudvig, R.; Fisher, J.; Jadhav, A. Integrated management of water hyacinth in South Africa. *WRC Rep.* **2010**.

44. Owens, C.S.; Madsen, J.D. Low temperature limits of waterhyacinth. *J. Aquat. Plant Manag.* **1995**, *33*, 63–68.

45. Miskella, J.J.; Madsen, J.D. The effect of temperature on waterhyacinth stem base regrowth 2019.

46. Lawrence, A.; Hoffmann, S.; Beierkuhnlein, C. Topographic diversity as an indicator for resilience of terrestrial protected areas against climate change. *Glob. Ecol. Conserv.* **2020**, e01445.

47. VonBank, J.A.; Casper, A.F.; Pendleton, J.E.; Hagy, H.M. Water hyacinth (Eichhornia crassipes) invasion and establishment in a temperate river system. *River Res. Appl.* **2018**, *34*, 1237–1243.

48. Hill, M.P. The impact and control of alien aquatic vegetation in South African aquatic ecosystems. *African J. Aquat. Sci.* **2003**, *28*, 19–24.

49. Agami, M.; Reddy, K.R. Competition for space between Eichhornia crassipes (Mart.) Solms and Pistia stratiotes L. cultured in nutrient-enriched water. *Aquat. Bot.* **1990**, *38*, 195–208.

50. Gopal, B. *Water hyacinth.*; Elsevier Science Publishers, 1987; ISBN 0444427066.

51. Muramoto, S.; Aoyama, I.; Oki, Y. Effect of salinity on the concentration of some elements in water hyacinth (Eichhornia crassipes) at critical levels. *J. Environ. Sci. Heal. Part A* **1991**, *26*, 205–215.

52. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28.

53. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

54. 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 (Cop.).* **2013**, *36*, 27–46.

55. Gomes, L.C.; Faria, R.M.; de Souza, E.; Veloso, G.V.; Schaefer, C.E.G.R.; Fernandes Filho, E.I. Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma* **2019**, *340*, 337–350.

56. Pullanagari, R.R.; Kereszturi, G.; Yule, I. Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. *Remote Sens.* **2018**, *10*, 1117.

57. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **2020**, *11*, 1–11.

58. Valavi, R.; Elith, J.; Lahoz‐Monfort, J.J.; Guillera‐Arroita, G. block CV: An r package for generating spatially or environmentally separated folds for k‐fold cross‐validation of species distribution models. *Methods Ecol. Evol.* **2019**, *10*, 225–232.

59. Meyer, H.; Reudenbach, C.; Wöllauer, S.; Nauss, T. Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecol. Modell.* **2019**, *411*, 108815.

60. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

61. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, 6.

62. Allen, G.H.; Pavelsky, T.M. Global extent of rivers and streams. *Science (80-. ).* **2018**, *361*, 585–588.

63. Theobald, D.M.; Harrison-Atlas, D.; Monahan, W.B.; Albano, C.M. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PLoS One* **2015**, *10*, e0143619.

64. Schulze, R.E.; Maharaj, M. Median First and Last Dates of Heavy Frost, their Variability, and the Duration of the Frost Period. *Schulze, RE* **2007**.

65. Dottori, F.; Alfieri, L.; Salamon, P.; Bianchi, A.; Feyen, L.; Lorini, V. Flood hazard map for Europe, 200-year return period. *Eur. Comm. Jt. Res. Cent.* **2016**.

66. Hengl, T.; Miller, M.A.E.; Križan, J.; Shepherd, K.D.; Sila, A.; Kilibarda, M.; Antonijević, O.; Glušica, L.; Dobermann, A.; Haefele, S.M. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Sci. Rep.* **2021**, *11*, 1–18.

67. Wu, Q. geemap: A Python package for interactive mapping with Google Earth Engine. *J. Open Source Softw.* **2020**, *5*, 2305.

68. Ali, M. PyCaret: An open source, low-code machine learning library in Python. *PyCaret version* **2020**, *2*.

69. Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D.D. Hyperopt: a python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **2015**, *8*, 14008.

70. Barrett, P.; Hunter, J.; Miller, J.T.; Hsu, J.-C.; Greenfield, P. matplotlib--A Portable Python Plotting Package. In Proceedings of the Astronomical data analysis software and systems XIV; 2005; Vol. 347, p. 91.

71. Waskom, M.L. Seaborn: statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021.

72. Arribas-Bel, D. Contextily 2021.

73. Gettys, L.A.; Haller, W.T.; Petty, D.G. Biology and control of aquatic plants. *A best Manag. Pract. handbook, 3rd edn. Aquat. Ecosyst. Restor. Found. Marietta* **2014**.

74. Madsen, J.D.; Luu, K.T.; Getsinger, K.D. *Allocation of Biomass and Carbohydrates in Waterhyacinth (Eichhornia crassipes): Pond-Scale Verification.*; ARMY ENGINEER WATERWAYS EXPERIMENT STATION VICKSBURG MS ENVIRONMENTAL LAB, 1993;

75. Hill, M.P.; Cilliers, C.J. A review of the arthropod natural enemies, and factors that influence their efficacy, in the biological control of water hyacinth, Eichhornia crassipes (Mart.) Solms-Laubach (Pontederiaceae), in South Africa. *A Rev. arthropod Nat. enemies, factors that Influ. their Effic. Biol. Control water hyacinth, Eichhornia crassipes Solms-Laubach (Pontederiaceae), South Africa.* **1999**, 103–112.

76. Hill, M.P.; Olckers, T. Biological control initiatives against water hyacinth in South Africa: constraining factors, success and new courses of action. In Proceedings of the ACIAR proceedings; ACIAR; 1998, 2000; pp. 33–38.

77. Venter, N.; Cowie, B.W.; Witkowski, E.T.F.; Snow, G.C.; Byrne, M.J. The amphibious invader: Rooted water hyacinth’s morphological and physiological strategy to survive stranding and drought events. *Aquat. Bot.* **2017**, *143*, 41–48.

78. Coetzee, J.A.; Center, T.D.; Byrne, M.J.; Hill, M.P. Impact of the biocontrol agent Eccritotarsus catarinensis, a sap-feeding mirid, on the competitive performance of waterhyacinth, Eichhornia crassipes. *Biol. Control* **2005**, *32*, 90–96.

79. Strange, E.F.; Landi, P.; Hill, J.M.; Coetzee, J.A. Modeling top-down and bottom-up drivers of a regime shift in invasive aquatic plant stable states. *Front. Plant Sci.* **2019**, *10*, 889.

80. Center, T.D.; Spencer, N.R. The phenology and growth of water hyacinth (Eichhornia crassipes (Mart.) Solms) in a eutrophic north-central Florida lake. *Aquat. Bot.* **1981**, *10*, 1–32.

81. Pérez, E.A.; Coetzee, J.A.; Tellez, T.R.; Hill, M.P. A first report of water hyacinth (Eichhornia crassipes) soil seed banks in South Africa. *South African J. Bot.* **2011**, *77*, 795–800.

82. Cilliers, C.J. Biological control of water hyacinth, Eichhornia crassipes (Pontederiaceae), in South Africa. *Agric. Ecosyst. Environ.* **1991**, *37*, 207–217.

83. Neiff, J.J.; Poi, A.S.G.; Casco, S.L. The effect of prolonged floods on Eichhornia crassipes growth in Paraná River floodplain lakes. **2001**.

84. Guillera‐Arroita, G. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography (Cop.).* **2017**, *40*, 281–295.

85. Normand, S.; Ricklefs, R.E.; Skov, F.; Bladt, J.; Tackenberg, O.; Svenning, J.-C. Postglacial migration supplements climate in determining plant species ranges in Europe. *Proc. R. Soc. B Biol. Sci.* **2011**, *278*, 3644–3653.

86. Mothilal, R.K.; Sharma, A.; Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020; pp. 607–617.

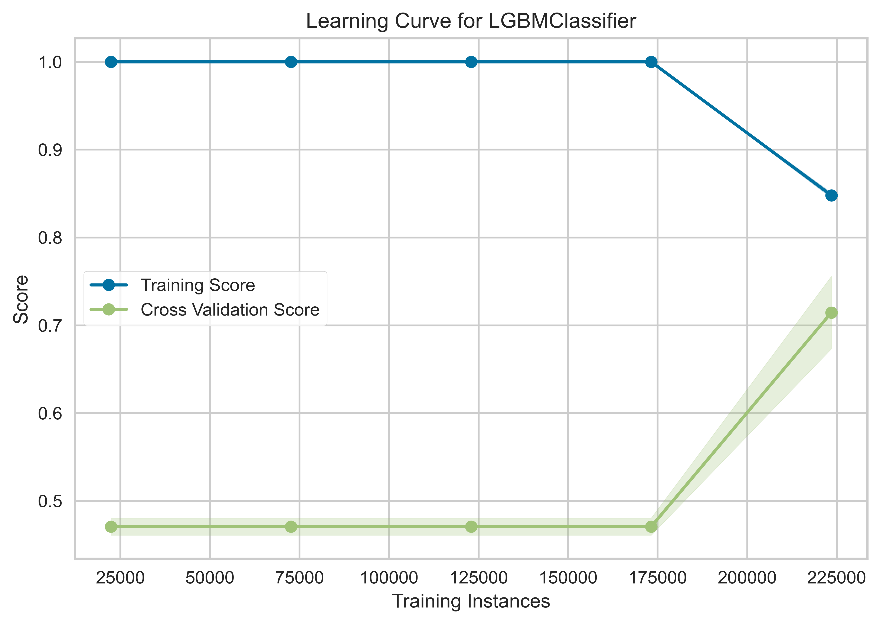
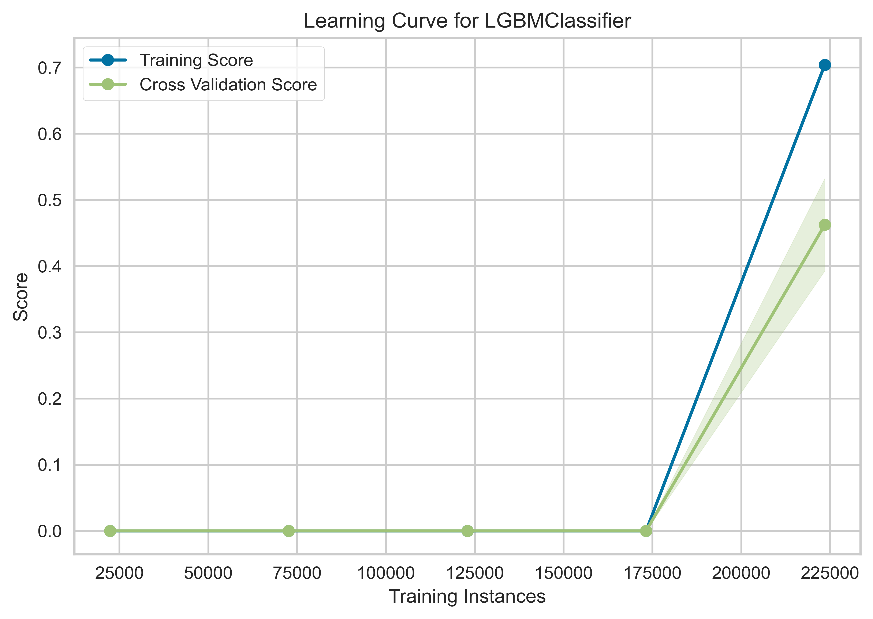
87. Yu, K.; Guo, X.; Liu, L.; Li, J.; Wang, H.; Ling, Z.; Wu, X. Causality-based feature selection: Methods and evaluations. *ACM Comput. Surv.* **2020**, *53*, 1–36.

1. A picture containing chart

   Description automatically generatedA picture containing diagram

   Description automatically generated**Supplementary information**

Figure S1: The differences in SHAP values and variable importance with block cross-validation (bottom) and without block cross-validation (top).



MCC

F1

Figure S2: Learning curve showing MCC and F1 score with the number of training instances. Both curves suggest that the lightGBM model requires more training data to converge.

These results are surprising. I expected the train-test accuracy to converge and with much less data. Previous studies that considered the effect of dataset size on accuracy for SDMs showed convergence for much smaller dataset sizes (~100 observations, refer to Stockwell & Peterson, 2002). However, the Stockwell paper did not use any sort of spatial cross validation. More recent studies that do use spatial cross-validation tend to use many more observations (for example Helmstetter et al., 2020 used ~ 8000 and achieved an AUC between 0.7 and 0.85). However, they do not speak about model convergence.

I am wondering if there are some factors specific to invasive species that could explain why the model has not converged, despite this large dataset. If I do end up testing this further, I could drop the absence data and use maxent with the presence only data for the comparison.

Figure S3: SHAP dependence plots for the remaining 14 predictors (in order of variable importance, variables 4-7) of water hyacinth with variable interactions (left) and main interactions (right).

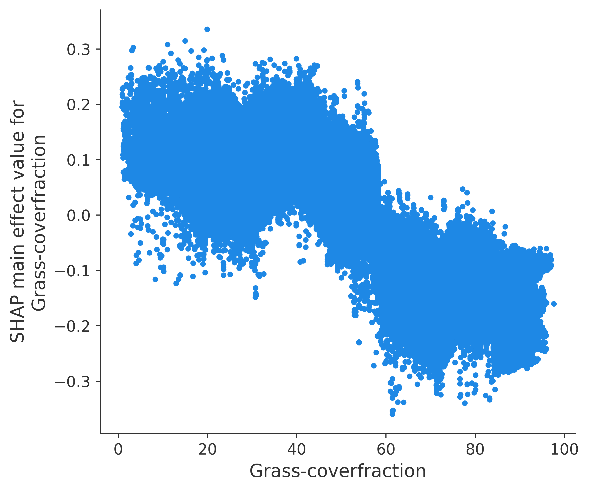
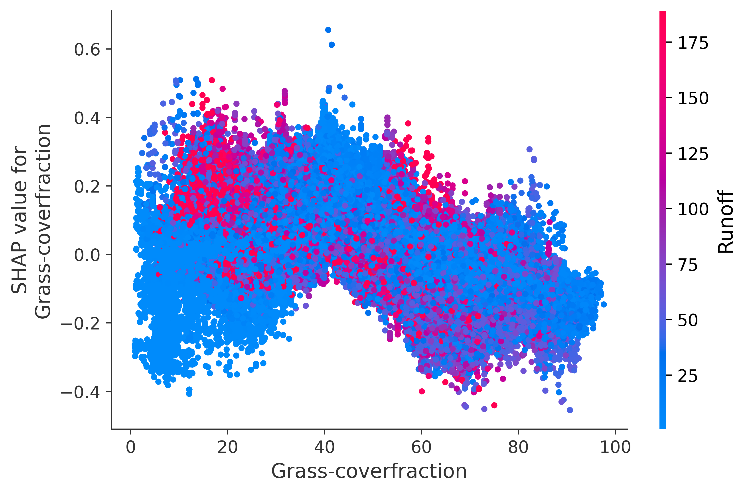
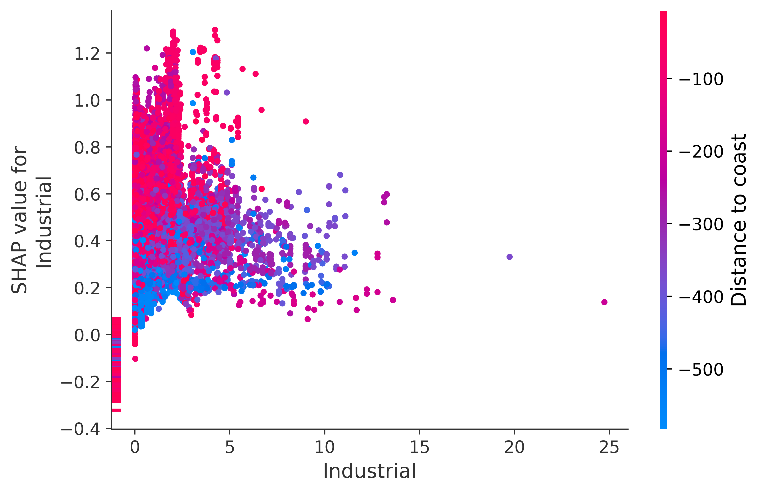
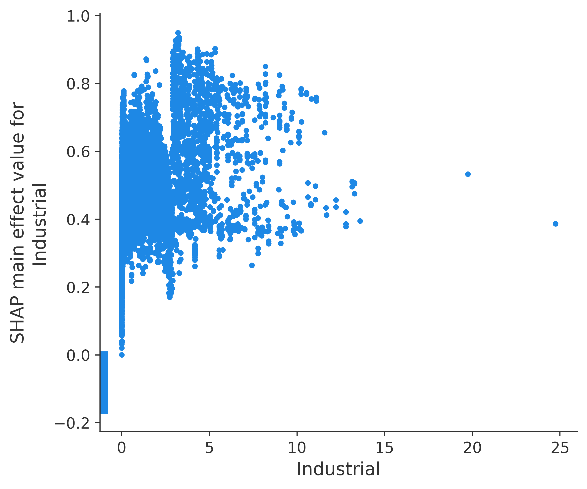
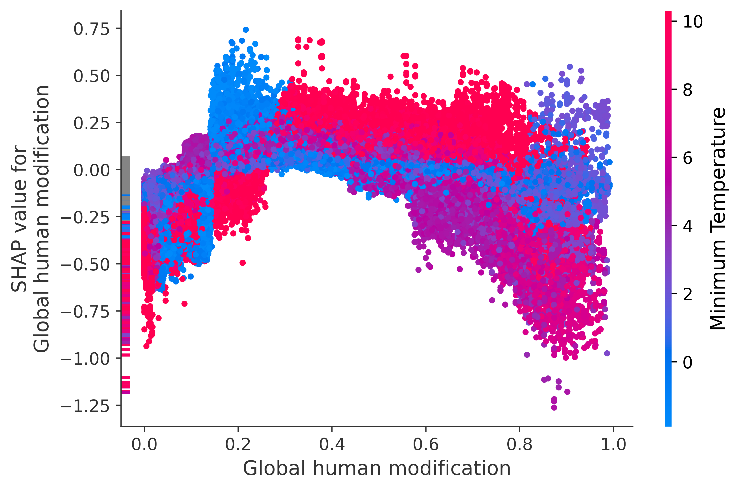
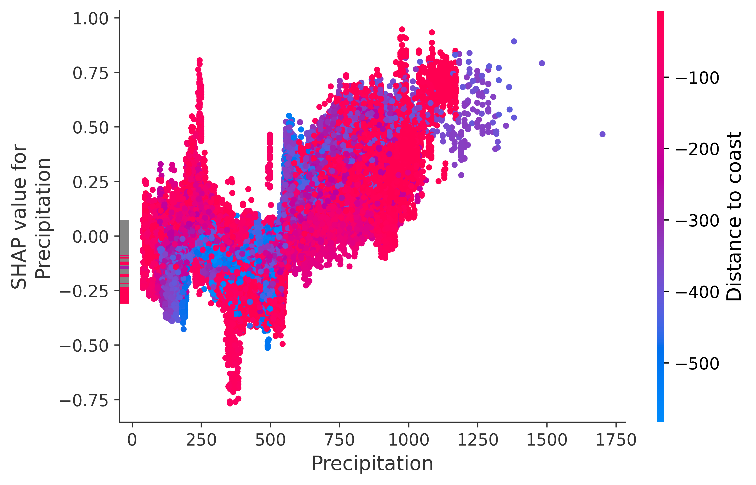
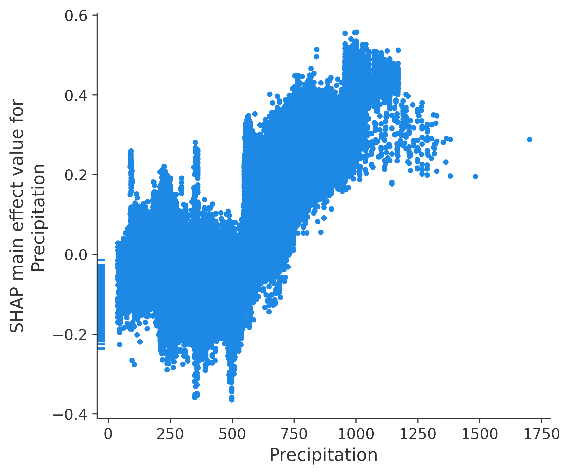


Figure S3 (continued): SHAP dependence plots for the remaining 14 predictors (in order of variable importance, variables 8-11) of water hyacinth with variable interactions (left) and main interactions (right).

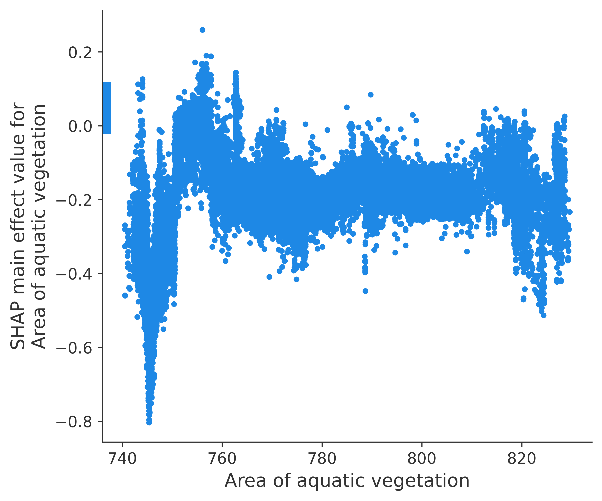
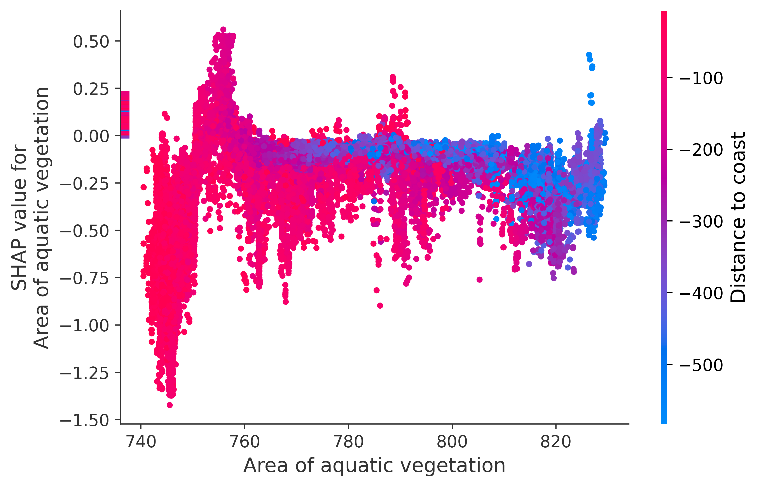
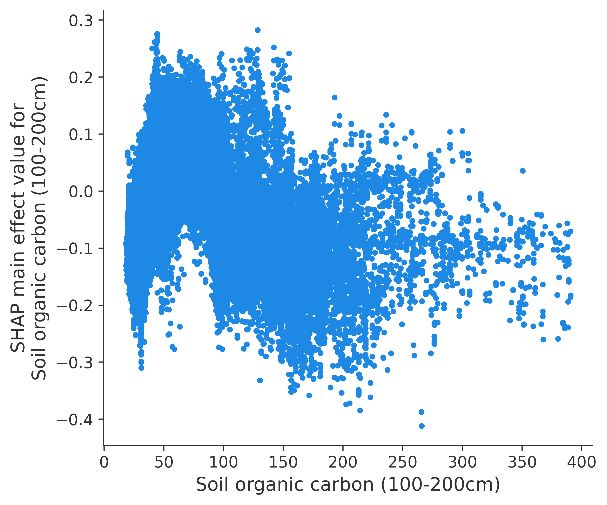
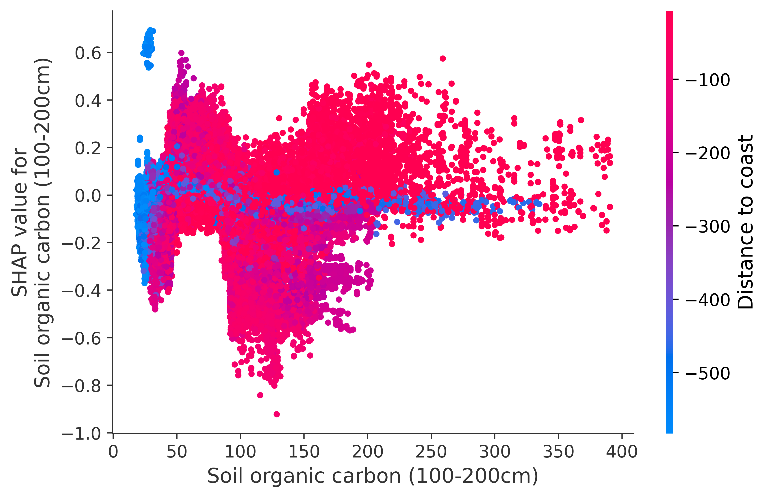
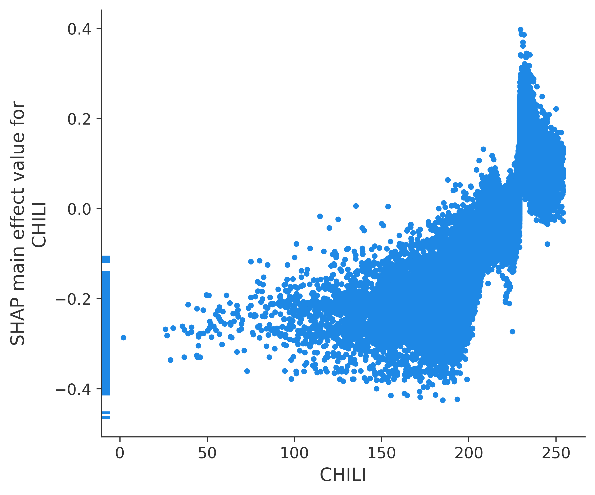
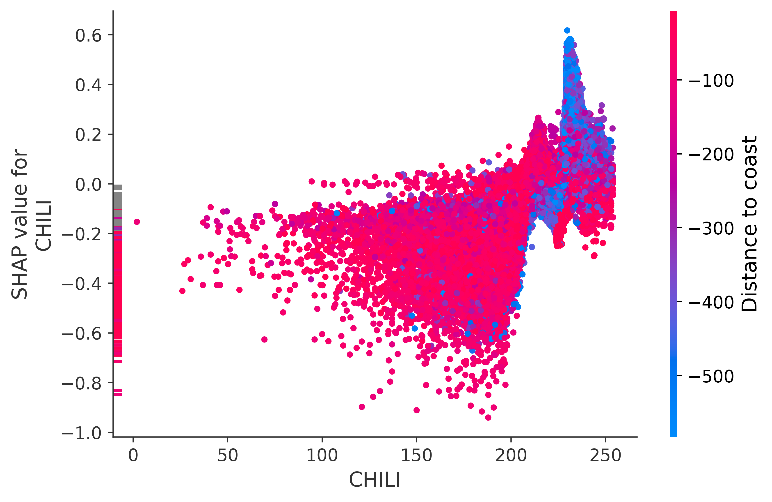
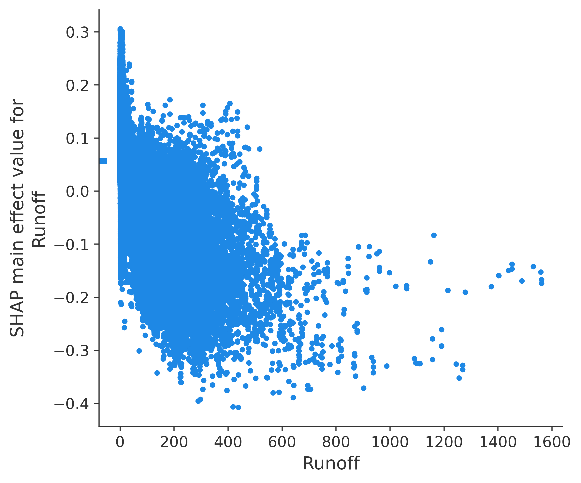


Figure S3 (continued): SHAP dependence plots for the remaining 14 predictors (in order of variable importance, variables 12-15) of water hyacinth with variable interactions (left) and main interactions (right).

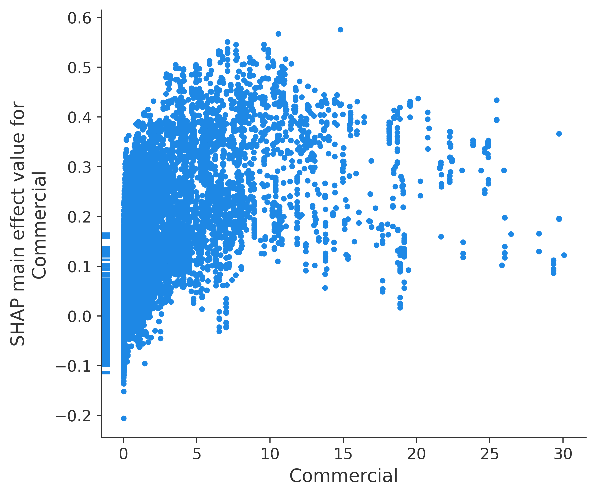
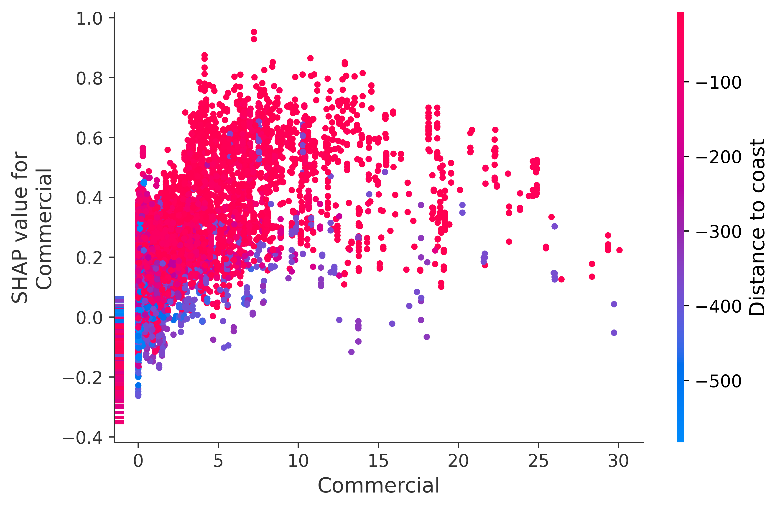
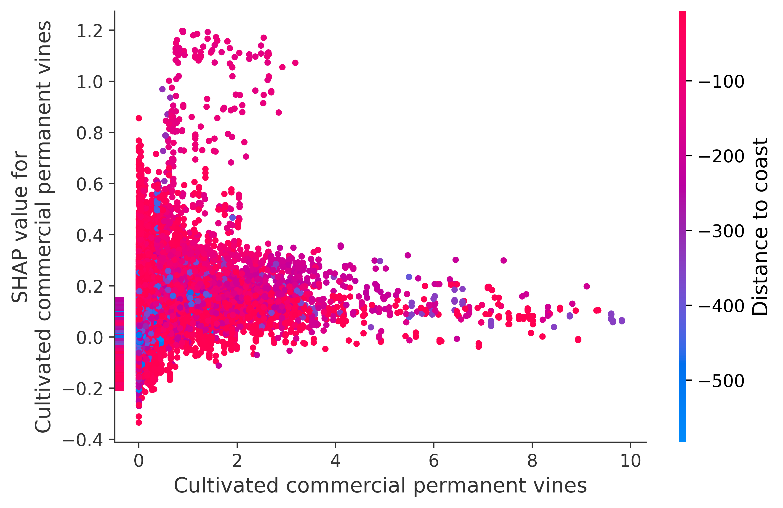
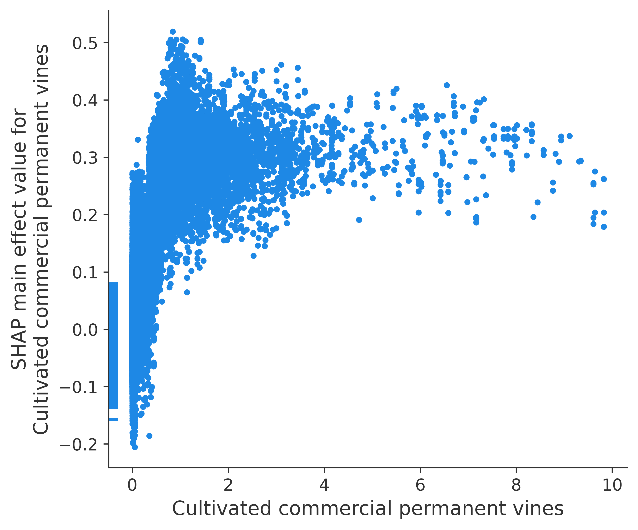
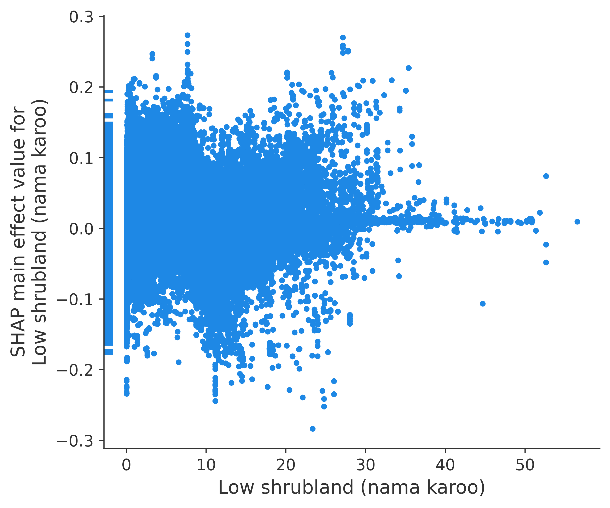
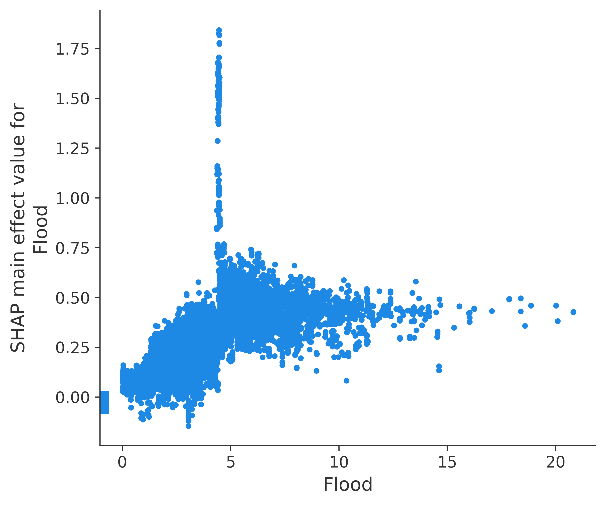


Figure S3 (continued): SHAP dependence plots for the remaining 14 predictors (in order of variable importance, variables 16 & 17) of water hyacinth with variable interactions (left) and main interactions (right).

