Spatial Analysis of Within-Lake Vulnerability to Aquatic Invasive Species

May 2023

PRESENTED TO

The Nature Conservancy
Brian Greene, Zachary Simek, and Tammara Van Ryn

PRESENTED BY

Tetra Tech
Kateri Salk, Brian Pickard, Mark Fernandez

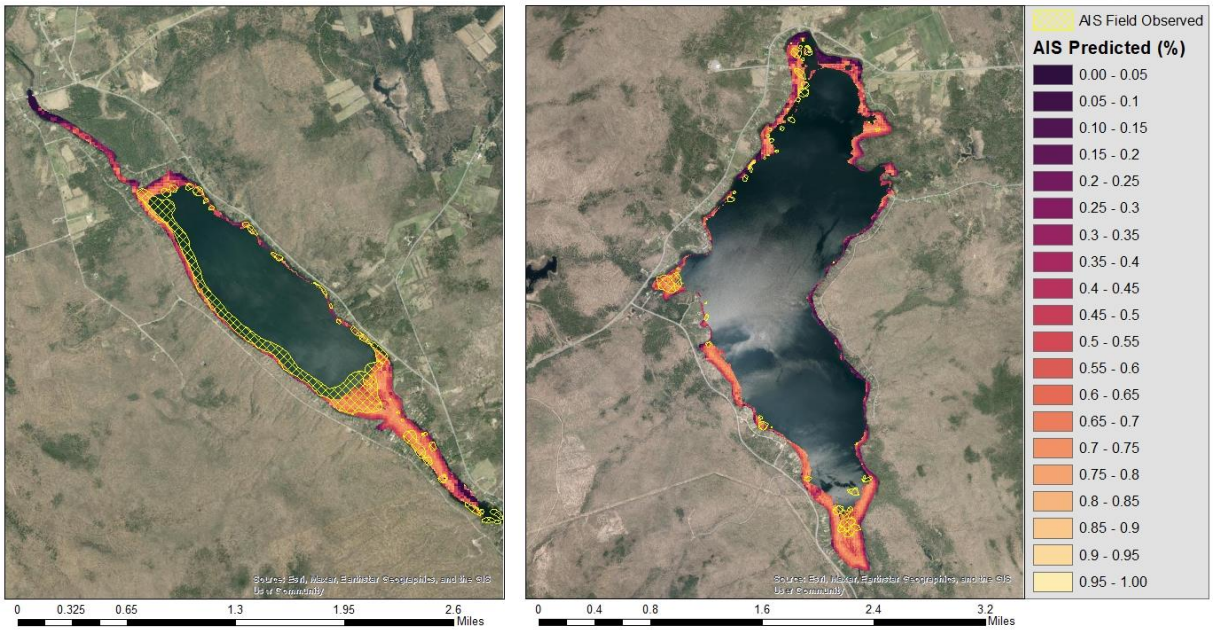
TETRA TECH

EXECUTIVE SUMMARY

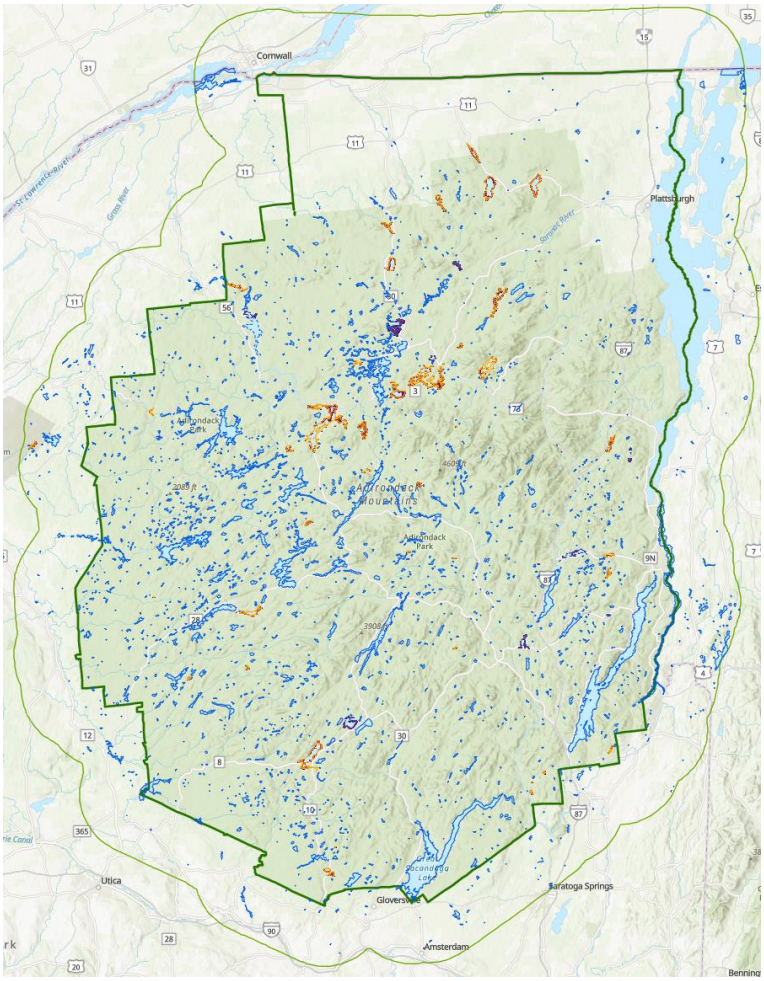
Aquatic invasive species (AIS) have significant negative impacts on lake ecosystems, underscoring the need for improved detection and management. The Adirondack Park Invasive Plant Program (APIPP) is dedicated to minimizing the impacts of invasive species in the Adirondack region, including those found in lakes. However, on-the-ground monitoring efforts conducted by APIPP are time- and resource-intensive. Statistical models could optimize monitoring efforts by targeting areas of likely AIS presence both within and across lakes. Recent advancements in satellite-based technology and machine learning algorithms present a promising pathway to model and predict intra-lake characteristics and likelihood of invasion.

The goals of this project were to train a statistical model to predict AIS presence within lakes in the Adirondack Partnership for Regional Invasive Species Management (PRISM) and to develop a web-based interactive map displaying areas of likely AIS presence or vulnerability. To achieve these goals, a large spatial dataset was created that included areas of mapped AIS locations, sonar-derived lake conditions (depth, hardness, and plant biovolume), proximity to anthropogenic points of interest (e.g., boat launches, beaches, campsites), and adjacent land cover. Then, three statistical models were tested head-to-head to determine the model that best predicted AIS presence. These models included a linear regression, a tree-based machine learning model, and an artificial neural network. The machine learning model (XGBoost) had the best model performance, correctly predicting AIS presence in ¾ of locations within the lakes with known AIS coverage. The most important predictor variables were proximity to plant biovolume, shoreline, forested land cover, impervious land cover, and agricultural land cover. However, the inclusion of biovolume as a predictor variable provided only a marginal increase in model performance while representing a large investment to acquire sonar data. Therefore, the final model predictor variables included only proximity to shoreline, forest, impervious cover, and agriculture, which provided the added benefit of national-scale data availability.

The model revealed that AIS cover was most likely in areas close to the shoreline (0-200 m), in areas 200 and 500-750 m from forested land cover, in areas with 0% or ~50% impervious cover in the immediately adjacent zone, and in areas with 0% and 10% agricultural cover in the immediately adjacent zone. Predicted AIS probability was not linear across the range of a given predictor variable, highlighting the usefulness of using machine learning models to characterize the relationships. Moreover, each 10x10 m “pixel” within a lake was assigned a probability of AIS presence, enabling mapping of AIS likelihood and vulnerability across the study lakes to inform monitoring efforts.



The model was applied to thousands of lakes in the Adirondack PRISM, enabling the creation of a predictive gridded heat map that displays areas of likely AIS coverage or vulnerability to future invasion in areas that have not been monitored previously. The map is published online as an interactive tool to enable APIPP, monitoring groups, stakeholders, and the public to explore AIS predictions in lakes across the Adirondack region.



The development of models to predict locations within lakes that are likely areas of AIS or are vulnerable to future AIS invasion represents a novel advancement at the intersection of statistical modeling, environmental big data, and lake management. Results of this work will enable prioritization of monitoring efforts for early detection surveys and other mitigation measures as well as serve as an invasive species communication tool for stakeholders and the public in the Adirondack region.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ACRONYMS/ABBREVIATIONS

Acronyms	Definition
ADK	Abbreviation for Adirondack
AIC	Akaike Information Criterion
AIS	Aquatic Invasive Species
APIPP	Adirondack Park Invasive Plant Program
AUC	Area Under the Curve
CART	Classification and Regression Trees
LASSO	Least Absolute Shrinkage and Selection Operator
NLCD	National Land Cover Database
NY DEC	New York Department of Environmental Conservation
PCA	Principal Component Analysis
PRISM	Partnership for Regional Invasive Species Management
TNC	The Nature Conservancy

1.0 INTRODUCTION

The Adirondack Park Invasive Plant Program (APIPP), a program of The Nature Conservancy (TNC), is one of eight Regional Invasive Species Management Partnerships across New York and is chiefly concerned with protecting the Adirondack region from negative impacts of invasive species. To achieve this, APIPP collects a range of spatial data on the presence, location, and percent cover of aquatic invasive species (AIS) in Adirondack Lakes, as well as other data including sonar and other manmade features. Presently, APIPP estimates that approximately 25% of the 100 sampled Adirondack Lakes have AIS presence. Previous efforts linked landscape variables to predict which lakes are most vulnerable to AIS invasion (Shaker et al. 2017). This work found that lakes most vulnerable to AIS invasion were lakes with more highly developed catchments, those located nearer to other invaded lakes, and those associated with recreation activities such as game fishing. Alternatively, lakes associated with less AIS were those at higher elevation, those with forested catchments.

Although advancements have been made to understand the factors dictating vulnerability and invasion patterns *across* lakes, understanding the spatial heterogeneity and predicting invasion patterns *within* lakes remains a challenge. On-the-ground monitoring can be costly in terms of time and resources, so optimizing monitoring time by targeting likely vulnerable areas of lakes could be beneficial for lake management programs. Recent advancements in satellite-based sensors, spatial and temporal resolution of remote sensing tools, and machine learning algorithms present a promising pathway to model and predict intra-lake characteristics and likelihood of invasion. The development of models to predict locations within lakes that are likely areas of AIS or are vulnerable to future AIS invasion represents a novel advancement at the intersection of statistical modeling, environmental big data, and lake management. Moreover, results of this work will enable prioritization of monitoring efforts for early detection surveys and other mitigation measures.

The overarching goal of this work was to develop accurate predictions of AIS across lakes in Adirondack Park. Objectives of this work were to:

1. Acquire and compile data sources that are potentially predictive of AIS
2. Develop and calibrate a modeling framework to accurately predict AIS cover within lakes where AIS coverage is known
3. Apply the predictive model to a broader suite of lakes to predict areas of AIS establishment or vulnerability where AIS coverage is unknown
4. Create an intuitive, easy to follow graphical user interface enabling users to explore gridded heat maps depicting AIS invasion likelihood and areas of mapped AIS coverage

2.0 METHODS

2.1 DATA COMPILATION

Data were acquired from various sources, comprising parameters that were evaluated by the research team as potential predictors of AIS habitat suitability. Data sources are described in Table 1. The goals of this project were distinct from determining how likely a given lake is to be invaded, which has been addressed in previous work (Shaker et al. 2017). Therefore, variables were targeted as those that corresponded to individual locations within lakes rather than those that corresponded to the lake as a whole. Some landscape variables, such as land use variables, were made more spatially explicit by incorporating “distance to” metrics and “focal” metrics of density near each location. These metrics also helped to account for spatial autocorrelation in the dataset, as models without spatial structure assume that each location is independent of other adjacent locations. The spatial resolution of each location (i.e., pixel size) was selected as 10x10 m; this decision was driven by the desire to be as spatially explicit as possible while considering the maximum resolution of each of the datasets and minimizing downsampling error. Relevant spatial resolutions from the datasets that contributed to this decision were National Hydrography Dataset data at 30 m resolution and BioBase data at 5 m resolution. Data were processed performed using ArcPRO 3.1 and Python 3.8.

AIS presence and absence data for 73 species were collected from New York State’s iMapInvasives data layer, aided by a custom download by the New York Natural Heritage Program given the large size of the dataset. From the initial list of species, animals of interest were restricted to completely or mostly sessile organisms (e.g., mussels), leaving out organisms whose spatial location varies substantially (e.g., waterfleas, crayfishes, turtles, fish). APIPP provided further feedback on species of interest for management purposes (n = 16 species, four of which were designated as motile). The species of interest were divided into four major groupings: overall AIS (n = 12 species), AIS plants (n = 8 species), AIS animals (n = 4 species), and AIS plants with high data richness (n = 5 species). Modeling was conducted primarily on the AIS plants with high data richness (i.e., Eurasian Watermilfoil, Curly Pondweed, European Frogbit/Common Frogbit, Variable Watermilfoil/Broadlead Watermilfoil, and Water Chestnut) because the high data density is likely to lead to better model output and these taxa are most likely to be easily identified in future monitoring efforts. Land cover metrics, generated from the National Land Cover Database (NLCD), were aggregated as defined in Table 2. Additional details on the points of interest are provided in the Appendix.

Table 1. Data sources, variables, and descriptions for lake metadata, AIS, and in-lake and landscape predictors.

Data Type	Data Source	Variables	Notes
Lake metadata	APIPP	Lake Name County Year of survey Binary presence/absence of AIS Binary presence/absence of BioBase data	Lake-scale data Surveys conducted 2018-2021 166 lakes had BioBase data

Data Type	Data Source	Variables	Notes
AIS data	NY State iMapInvasives	AIS presence/absence AIS species	Contained point, line, and polygon features. Point and line data were converted to polygons by assigning a buffer of 30 m. In the event of multiple sampling events, the most recent observation was selected.
Sonar-derived data	BioBase	Depth Hardness Biovolume	
Points of interest	NY DEC ADK Atlas APIPP	Group 1: Distance to impervious anthropogenic sources Group 2: Distance to camping Group 3: Distance to pervious anthropogenic sources Group 4: Distance to parking lot Group 5: Distance to fishing location Group 6: Distance to marina Group 7: Distance to hand boat launch Group 8: Distance to trailer boat launch Group 9: Distance to beach	Parameters had overlap between NY DEC and ADK Atlas. A single points of interest layer was created by combining these two data sources. Parameters were grouped based on common characteristics. Tetra Tech developed the groupings with APIPP guidance. Where polygons were specified, we maintained those extents. When a point was specified, a 15 m buffer was applied.
Adjacent landcover	NLCD	Distance to natural Distance to impervious Distance to agricultural Distance to wetland Distance to forest Density of natural cover in vicinity around pixel (120-m grid) Density of impervious cover in vicinity around pixel (120-m grid) Density of agricultural cover in vicinity around pixel (120-m grid) Density of wetland cover in vicinity around pixel (120-m grid) Density of forest cover in vicinity around pixel (120-m grid)	Hereafter called “Distance to” and “Focal” metrics
Miscellaneous	Calculated	Binary shoreline Elevation	

Table 2. Land cover classes calculated from NLCD.

Land Cover Metric	National Land Cover Classes
Natural	Barren Land/Deciduous Forest/Evergreen Forest/Mixed Forest/Shrub/Scrub/Herbaceuous/Woody Wetlands/Emergent Herbaceuous Wetlands
Impervious	Developed, Open Space/Developed, Low Intensity/Developed, Medium Intensity/Developed, High Intensity
Agricultural	Hay/Pasture/Cultivated Crops
Wetland	Woody Wetlands/Emergent Herbaceuous Wetlands
Forest	Deciduous Forest/Evergreen Forest/Mixed Forest

Macrophyte biovolume was incorporated as a potential predictor variable for AIS. Biovolume is inherently not independent of AIS, given that the AIS species of interest are macrophytes. However, exploratory analysis indicated that biovolume proportion cover and AIS presence did not have a strong relationship (Figure 1), which may suggest that areas within lakes that are conducive to macrophyte growth may be similar for both native and invasive species. Further, TNC was interested in evaluating whether collecting biovolume data is useful in determining lake vulnerability to AIS invasion. For these reasons, biovolume was retained as a predictor of interest but derived “distance to” and “focal” metrics for this variable to account for potential spatial autocorrelation between biovolume and AIS presence. The best fit model was run with and without the biovolume predictors to determine the importance of biovolume as a predictor; this may be a useful piece of information for TNC to inform future monitoring efforts and investment in BioBase data collection.

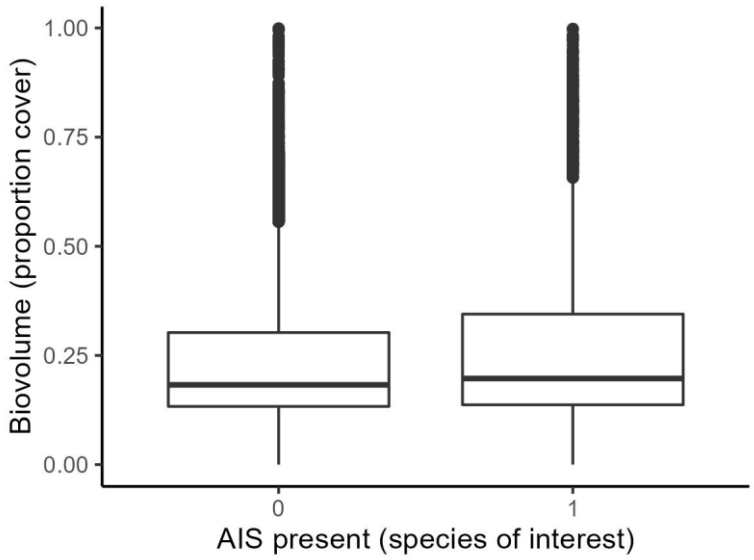


Figure 1. Biovolume in relation to AIS presence and absence.

The dataset includes data across entire lakes, which in some cases includes deeper depths that are not suitable for macrophyte growth. In order to provide the model with relevant sampling areas, an upper depth cutoff was explored. One spatially explicit option would be to use biovolume data to establish an occupancy depth for each lake. However, this may result in the model receiving variable maximum depths by lake, which would lead to difficulty in establishing the potential impact of depth on AIS presence. Rather, a uniform depth cutoff was generated for consistency across the dataset, with the caveat that this depth cutoff could be over- or underestimated for a given lake. To choose a depth cutoff, AIS presence/absence data were examined across depths. This pattern showed variable AIS coverage across depths, with AIS presence leveling off at a low value beyond 8 m depth (Figure 2). Therefore, a depth cutoff of 7 m (23 ft) was implemented as the upper bound for modeling.

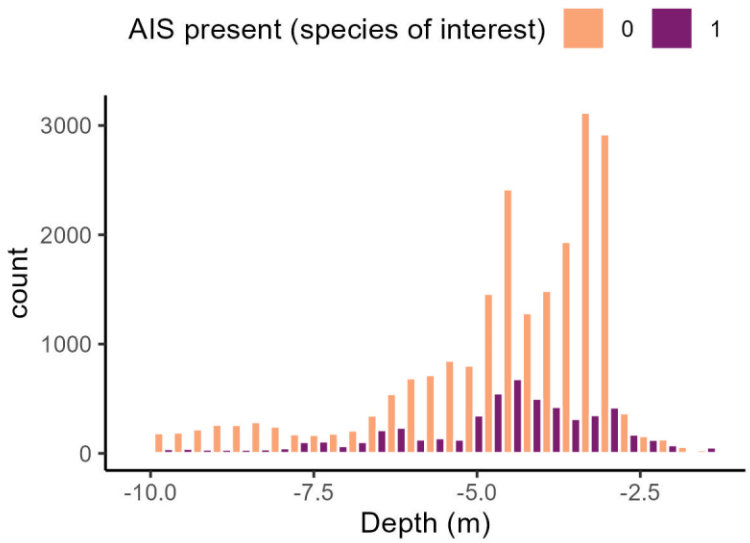


Figure 2. Water depths in relation to AIS presence and absence.

2.2 EXPLORATORY ANALYSIS

We evaluated correlations among predictor variables and dimension reduction analysis on the predictors to explore the relationships among predictors and potentially inform feature selection. This exploratory analysis was evaluating using correlation coefficients and principal component analysis (PCA). This analysis allowed for the evaluation of variables that covaried across the dataset as well as evaluate the multivariate distributions of observations that made up the model training, testing, and application groups to determine how well their distributions overlapped. Exploratory analysis was conducted in R (R Core Team 2022).

2.3 MODEL TRAINING, TESTING, AND SELECTION

The general modeling goal was to predict the presence or absence of AIS (binary response variable) using a combination of several predictor variables. The AIS dataset of interest was AIS plant species with high data richness (i.e., Eurasian Watermilfoil, Curly Pondweed, European Frogbit/Common Frogbit, Variable Watermilfoil/Broadleaf Watermilfoil, and Water Chestnut). Modeling was conducted using a tiered approach that tested the performance of three models against each other, ranging from linear models to complex machine learning approaches. The three models tested were a linear regression, a tree-based machine learning model, and an artificial neural network. These models vary in their assumptions, degree of complexity, ease of use, and interpretability. Comparing these models head-to-head allowed for evaluation of the best-performing model structure.

The linear model consisted of a logistic regression, which predicts a binary (i.e., presence/absence) response from a set of predictors. To enable model selection, we used LASSO (Least Absolute Shrinkage and Selection Operator; Tibshirani 1996) regression. LASSO is a shrinkage method, meaning it receives an inclusive list of predictor variables and selects a subset of those predictor values by shrinking the non-subsetted predictor coefficients to zero. To assign coefficient values, LASSO attempts to minimize the sum of squared residuals while penalizing for coefficients with larger magnitudes. This penalizing encourages coefficients that are smaller in magnitude, hence the term “shrinking.” LASSO presents several benefits over other methods. Stepwise selection methods such as forward or backward selection via Akaike Information Criterion (AIC) may not select the best combination of predictors if predictors are correlated. Ridge regression is similar to LASSO, though it does not shrink predictor coefficients to zero and is thus not a true model selection method.

Skewed predictor variables were square-root transformed to approximate a normal distribution, where appropriate. LASSO has one hyperparameter for tuning (the shrinkage parameter). LASSO was tuned across a grid of shrinkage values using 10-fold cross validation on the training dataset. The shrinkage value producing the minimum model deviance was identified. To account for random splits, this process was repeated 20 times. The mean of those 20 values was the final shrinkage value.

Decision tree-based machine learning models use a tree-like structure that sequentially selects features (predictor variables) that are predictive of the response, with each “node” of the tree producing a split that grows the tree at each subsequent level. Common tree-based models include random forest and classification and regression trees (CART). Gradient boosting (Friedman 2001) decision tree algorithms are similar to a CART model but place more importance on mis-classified observations, thereby attempting to concentrate model improvements on areas where the existing trees perform poorly. These algorithms have begun to dominate the machine learning space for their speed and predictive capabilities relative to other algorithms. A recent advance in this space has been the creation of GPBoost (Chen and Guestrin 2016), an algorithm that combines gradient boosting decision trees with Gaussian process models, which are often used to model spatial data. While GPBoost appeared to be an ideal model structure for this dataset, the algorithm is very new and has not been optimized for datasets larger than a few thousand data points. Given the limitations of processing time, we opted to use XGBoost, a similar gradient boosting method that does not use a Gaussian process.

The XGBoost model was calibrated, or hypertuned, on the training data to optimize model performance while avoiding overfitting. The following parameters were tuned during calibration. We used a deterministic grid search followed by k-fold cross validation that tuned the following parameters:

- Number of boosting iterations (trees)
- Learning rate
- Maximum tree depth
- Minimum samples per leaf
- Number of leaves
- Learning rate

Neural networks use a series of processors operating in parallel and arranged in layers. The first tier receives the predictor variables, and then the model creates successive tiers where the predictor data is manipulated and fed into the next successive tier. The last tier produces the response prediction (probability of AIS presence). Neural networks capitalize on all available predictor variables, unlike modeling methods that perform feature selection. To calibrate model parameters, a calibration algorithm was run iteratively to test all possible combinations of parameters and their values, with the result being the optimized model. Parameters tuned for the neural network included:

- Batch size
- Epochs

- Learning rate
- Momentum
- Neuron activation function
- Dropout percentage
- Dropout weight
- Number of neurons in hidden layer

The dataset was divided into training and testing portions using an 80:20 random split. Because pixels within a given lake may not necessarily be independent of one another, we grouped the training/testing data at the lake level (i.e., a given lake only appeared in the training or testing dataset, never both). The initial training/testing split was identified randomly and was refined to ensure representative assignment by location, sampling intensity, and size. Lakes that were substantially hydrologically connected (e.g., chains of lakes, “upper” and “lower” lake bays) were assigned into the same group. The overall prevalence of AIS presence in the dataset was low (8.6%), and this can cause issues for evaluating model fit (i.e., if the model always predicts AIS absence, it will be correct >90% of the time). Therefore, so we further sub-sampled the training/testing dataset so that a 50:50 presence:absence ratio was obtained.

In exploring AIS spatial distributions across lakes, we found that some lakes had detailed spatial mapping of AIS, most often as polygons. Some other lakes had the entire lake marked as having AIS cover or marked AIS coverage as points. This is a result of the AIS database containing data collected by TNC as well as citizen groups, the latter of which may submit non-spatially detailed AIS information. An example of Upper and Lower Chateaugay Lakes illustrates spatially detailed mapping of AIS, and an example of Indian Lake illustrates AIS marked as points (Figure 3). To further refine the dataset, a subset of lakes was identified as having spatially detailed mapping and were used in model training and testing. This refinement adjusted the number of lakes in the training/testing dataset from 58 lakes to 42 lakes. 33 lakes were designated as training lakes, and 9 lakes were designated as testing lakes (Figure 4). The number of pixels for each group was 53,682 for training and 19,408 for testing. The same set of training/testing lakes and pixels was used across all models to enable direct comparison.

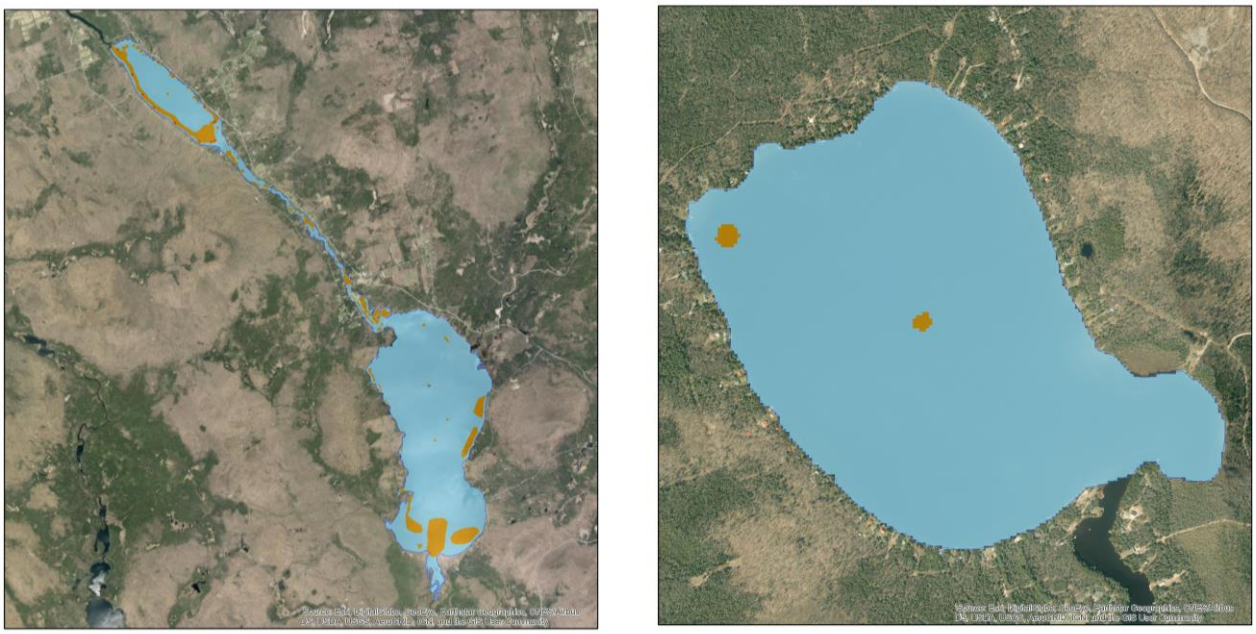


Figure 3. Examples of AIS marked as spatially detailed polygons (left; Upper and Lower Chateaugay Lakes) and as points (right; Indian Lake).

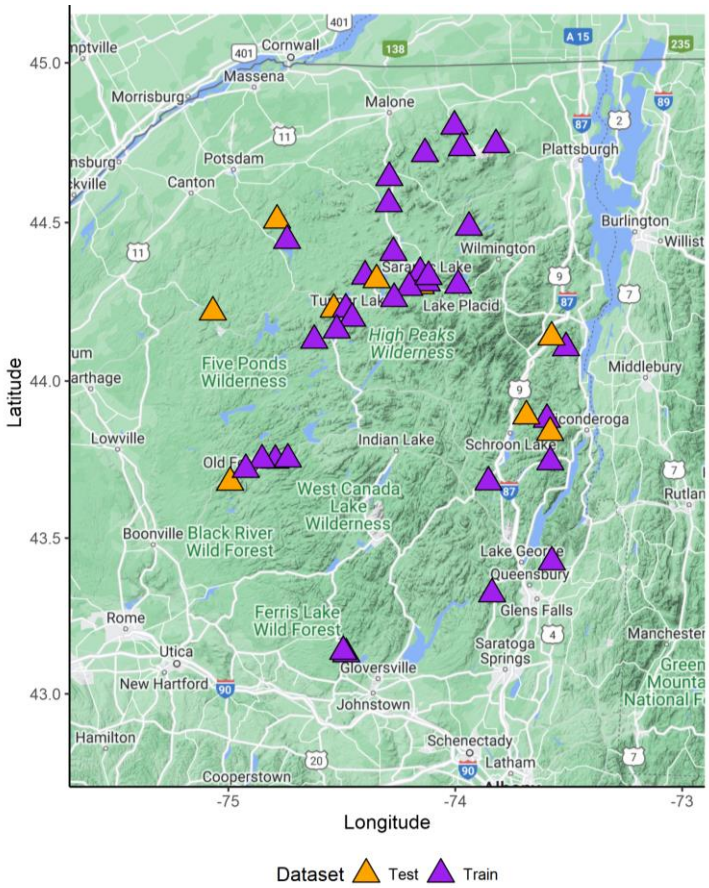


Figure 4. Map of model training and testing lakes.

Across all models, performance was lower than anticipated. Given the vetting and refinement of the predictor and response dataset, we ruled out that model performance could have been impacted by low quality input data. Another possible interpretation is that statistical models may become “confused” by predictor variables that have inconsistent and competing information. We pursued an additional model selection approach on the best-performing model called feature selection, which consisted of an algorithm that cycles through every possible subset of predictor variables and finds the number and combination of predictors that results in the best model fit.

Model output included a predicted probability of AIS presence for each pixel. Predicted probabilities >0.5 were assigned as present, and probabilities <0.5 were assigned as absence. The following performance metrics were evaluated across models:

- **Accuracy:** Ratio of correct predictions to total predictions. Used to identify overall performance of classification.
- **Precision:** Ratio of correct present (or absent) classifications to the total number of predicted positive (or absent) classifications. Used to identify the correctness of classification. Evaluated separately for presence and absence classifications.
- **Recall:** Ratio of correct present (or absent) classifications to the total number of present (or absent) classifications. Used to identify the sensitivity of classification. Evaluated separately for presence and absence classifications.
- **Area Under the ROC Curve (AUC):** aggregate measure of performance across all possible classification thresholds.

Each metric has a range from 0 to 1, with higher values indicating better model performance. Models were ranked based on the overall balance of model fit metrics.

Confusion matrices were also used to interpret classification output, showing the magnitudes of true positive and negative predictions as well as false positives and false negatives (Table 3). In addition to using confusion matrices to generate model fit metrics such as accuracy, precision, and recall, we examined these matrices to balance modeling goals with respect to the four quadrants. For instance, TNC monitoring efforts may desire to minimize false negatives, where the model predicts AIS absence but the true condition has AIS presence. A false negative in this case may cause monitoring groups to avoid monitoring areas that have AIS presence. A false positive, while it could cause wasted time and effort, would be a relatively better outcome than a false negative for AIS monitoring and management.

Table 3. Example confusion matrix.

	Predicted presence	Predicted absence
Observed presence	True positive	False negative (type 2 error)
Observed absence	False positive (type 1 error)	True positive

XGBoost and neural network modeling were performed using the “XGBoost” and “sklearn” packages in Python, respectively. Logistic regression was conducted using R (R Core Team 2022) and the “glmnet” package (Friedman et. al. 2010).

2.4 MODEL APPLICATION

Model application beyond the initial set of training/testing lakes was explored based on (a) the availability of predictor data, and (b) the similarity of lakes in the “apply” group compared to the training and testing lakes. To maximize the total number of lakes for which AIS cover could be predicted while maintaining geographic consistency, the area within Adirondack Park plus a 10-mile buffer around the park was selected. This area represents the Adirondack Partnership for Regional Invasive Species Management (PRISM) boundary. However, the models used in this analysis do not enable extrapolation, so the PRISM lake dataset was trimmed to avoid extrapolating the model beyond the conditions for which it was trained. Lakes smaller than 5 acres were excluded from analysis, and the largest lakes included in the dataset were Lake George (120 km^2), Great Sacandaga Lake (101 km^2), and Cranberry Lake (28 km^2). These lakes are larger than the largest lakes in the training and testing datasets (up to 25 km^2) but were indicated as lakes of interest by TNC. Individual pixels for lakes in the apply group were restricted to those with predictor variable values that were within the upper and lower bounds of the predictor variable values in the training dataset, designated as the minimum and 99th percentile values (maximum was not used due to the presence of a few exceptionally high outliers). The model was applied to over 1,000 lakes in the Adirondack Park PRISM boundary (the total count varied depending on how connected lakes and bays were lumped or split in the National Hydrography Dataset).

3.0 RESULTS

3.1 EXPLORATORY ANALYSIS

Correlation plots and PCA showed that forest and natural land cover were highly correlated, focal and distance metrics for corresponding variables were generally correlated, the points of interest groups (numbered groups) were correlated, and elevation was correlated with land cover metrics (Figure 5, Figure 6). For correlated variables, we expected that if one variable rose in feature importance, the other variable would fall. The model selection algorithms explicitly account for correlated predictors in this manner, so there was no need to remove potential predictor variables *a priori*. The PCA graph showed predictor variables occupying all four quadrants of principal component space for the first two principal components (x and y axes). This output indicated that the combination of BioBase, land cover, and points of interest data account for different aspects of variability in lake data across the dataset, suggesting promise for finding important statistical relationships with AIS cover.

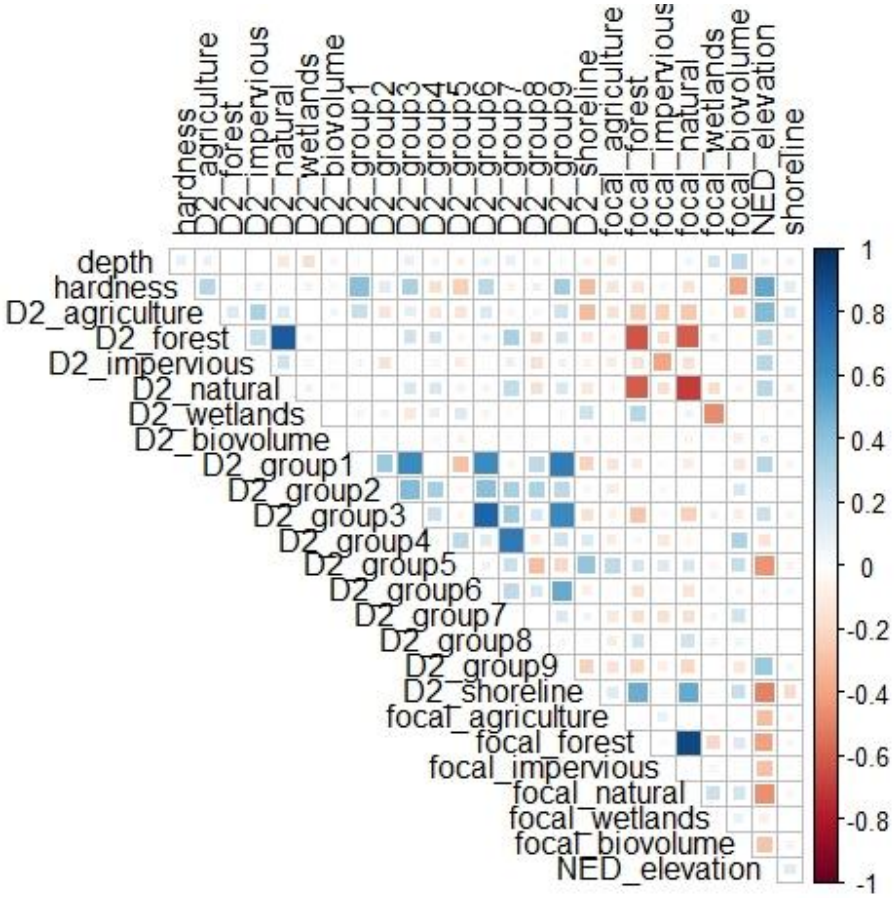


Figure 5. Pearson correlation coefficients among predictor variables.

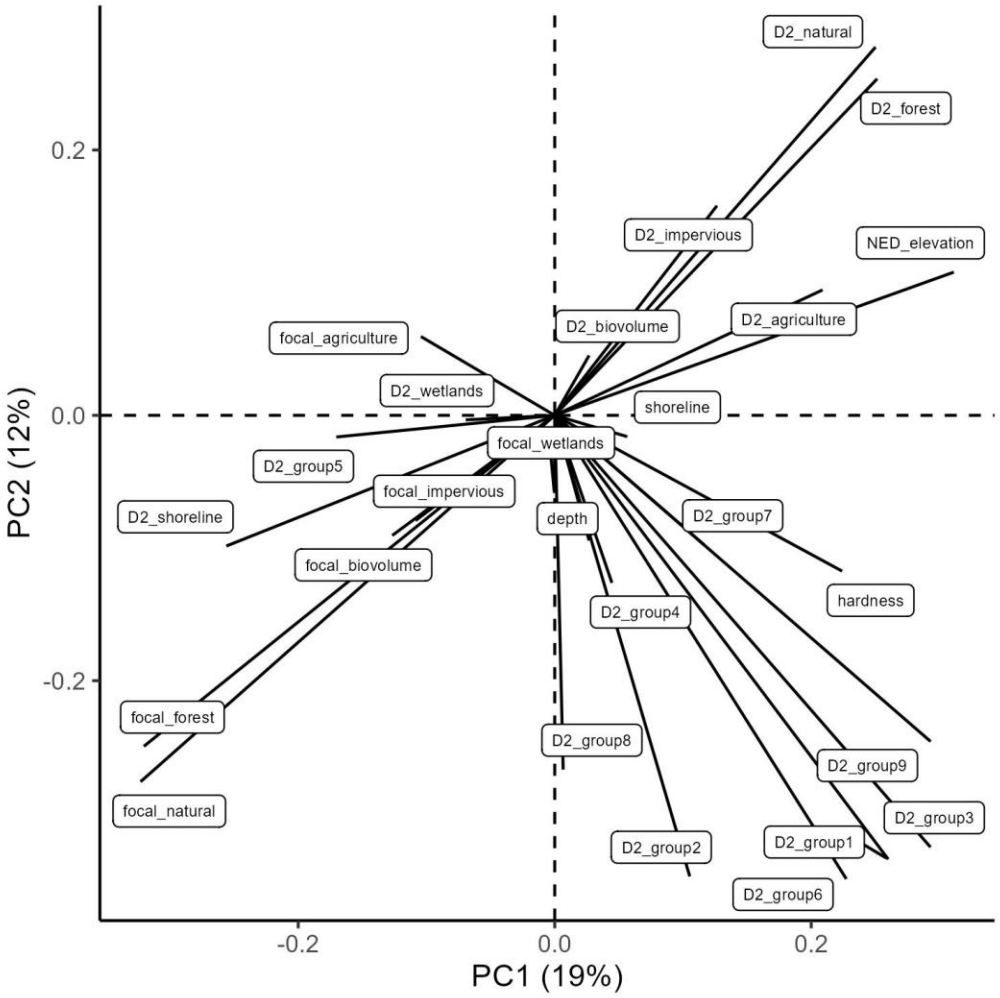


Figure 6. PCA with predictor variables shown as vectors. The first and second principal components explained 19 and 12% of variance in the dataset, respectively.

3.2 MODEL PERFORMANCE EVALUATION

The neural network performed worst, followed by the LASSO logistic regression model. The XGBoost model performed best, and model performance improved following the feature selection process. The “best subset” XGBoost model was ranked first in model performance, followed by the “best subset” XGBoost model that excluded the biovolume metrics as possible predictors (Table 4).

Table 4. Model performance metrics and rankings. The XGBoost “best subset” models had the best model performance and were explored head-to-head to determine the final model choice.

	Accuracy	Precision	Recall	AUC	Rank
Linear Model (LASSO)	0.540	0.54 *absent: 0.54* *present: 0.54*	0.55 *absent: 0.54* *present: 0.55*	0.617	4
Neural Network	0.446	0.40 *absent: 0.47* *present: 0.34*	0.45 *absent: 0.78* *present: 0.11*	0.366	5
XGBoost: Naïve run	0.565	0.63 *absent: 0.54* *present: 0.72*	0.57 *absent: 0.92* *present: 0.21*	0.707	3
XGBoost: Best subset	0.753	0.76 *absent: 0.81* *present: 0.72*	0.75 *absent: 0.73* *present: 0.77*	0.753	1*
XGBoost: Best subset w/o biovolume	0.708	0.71 *absent: 0.71* *present: 0.70*	0.71 *absent: 0.70* *present: 0.71*	0.789	1*

The XGBoost “best subset” model had the best model performance, and therefore the most correct predictions (Table 5). This model predicted 3,238 false negatives and about half as many false positives, 1,539.

Table 5. Confusion matrix for the XGBoost best subset model.

	Predicted presence	Predicted absence
Observed presence	6,466 (33%)	3,238 (17%)
Observed absence	1,539 (8%)	8,165 (42%)

The XGBoost “best subset without biovolume” model had the second best model performance, with fewer correct predictions than the “best subset” model (Table 6). However, this model had 312 fewer false negatives than the “best subset” model, while at the same time having more false positives.

Table 6. Confusion matrix for the XGBoost best subset model without biovolume.

	Predicted presence	Predicted absence
Observed presence	6,778 (35%)	2,926 (15%)
Observed absence	2,737 (14%)	6,967 (36%)

In addition, we examined the two XGBoost model subsets for their predicted probabilities. By default, the model defines a “presence” as any prediction >0.5 and an “absence” as any prediction <0.5. This plot showed that changing the threshold from 0.5 would not enhance model performance, since any adjustment to reduce false positives or negatives would lead to a corresponding flip from the true positive or true negative category into a false prediction (Figure 7).

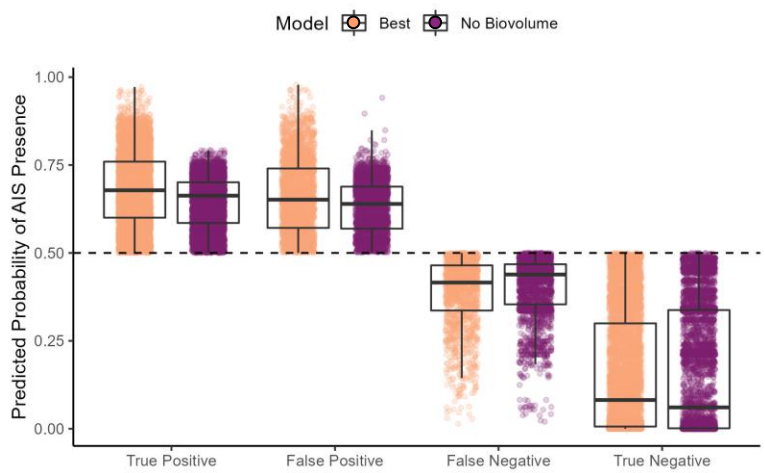


Figure 7. Distribution of predicted probabilities for the XGBoost best subset model and the XGBoost best subset model without biovolume.

3.3 VARIABLE IMPORTANCE

To provide insights about the importance of input variables, we examined the feature importance of variables selected by the models in more detail. Interpretation of predictor variables is more difficult for machine learning approaches such as XGBoost and neural networks than for linear regression approaches such as LASSO, because this “black box” model does not output model coefficients. Instead, the relative impact of individual variables in machine learning models can be evaluated using feature importance and partial dependence plots. The feature importance plots for the XGBoost best subset model (Figure 9) and the XGBoost best subset model without biovolume (Figure 8) show that distance to shoreline and distance to forest were both important predictors for the models. Focal density of biovolume was the most important predictor in the XGBoost best subset model, followed in order of importance by distance to shoreline, distance to forest, and distance to biovolume. For the XGBoost best subset model without biovolume, distance to shoreline was the most important predictor, followed in order of importance by distance to forest, focal density of impervious cover, and focal density of agricultural cover. Note that the F scores displayed are relative values intended to be compared in magnitude within an individual model rather than comparing values across models.

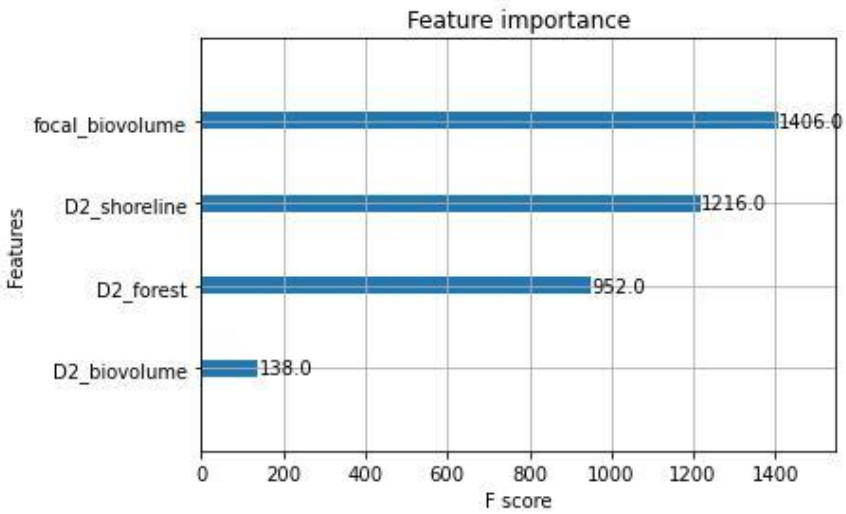


Figure 9. Feature importance for the XGBoost best subset model with biovolume. The F score represents the relative importance of each variable in the model.

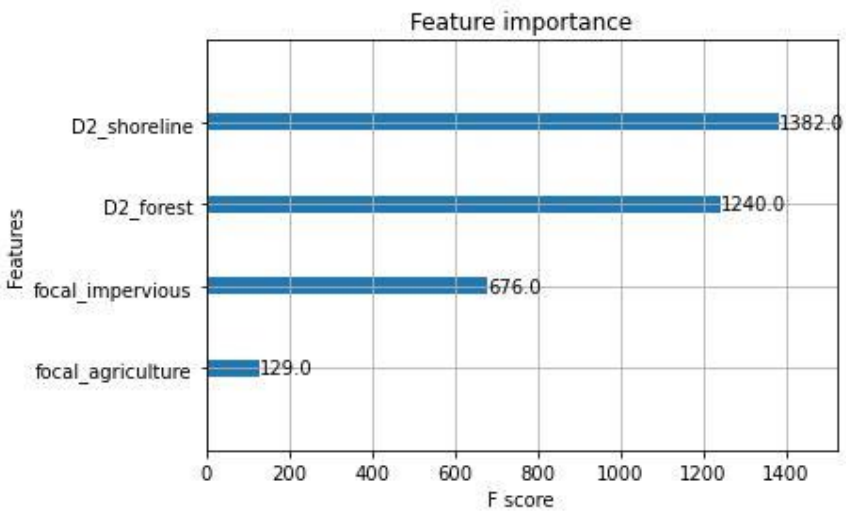


Figure 8. Feature importance for the XGBoost best subset model without biovolume. The F score represents the relative importance of each variable in the model.

Partial dependence plots show the predicted probability of AIS presence across a gradient of a given predictor variable. Plots for the variables in the XGBoost best subset model without biovolume are shown in Figure 10. Note that unlike the linear model, the probabilities across the range of a given variable need not be linear or directional. The partial dependence plot for distance to shoreline showed that the predicted probability of AIS was relatively high close to shore (0-200 m), with the predicted probability decreasing farther from shore (200-500 m). The predicted probability then rose again at 500 m from shore and farther, though these predicted probabilities were associated with a lower density of sampled points and may have been driven by a small number of large lakes in the training dataset. Predicted AIS presence displayed a multi-peaked relationship with distance to forest, with highest probabilities predicted at ~200 and 500-750 m and lowest probabilities predicted at 0 m and ~350 m. AIS presence had the highest predicted probabilities at 0 and 50% impervious cover in the focal area and the lowest predicted probabilities at 25% and >75% impervious cover. AIS had the highest predicted probability at 0% agricultural cover in the focal area and a decreased probability at >0% agricultural cover.

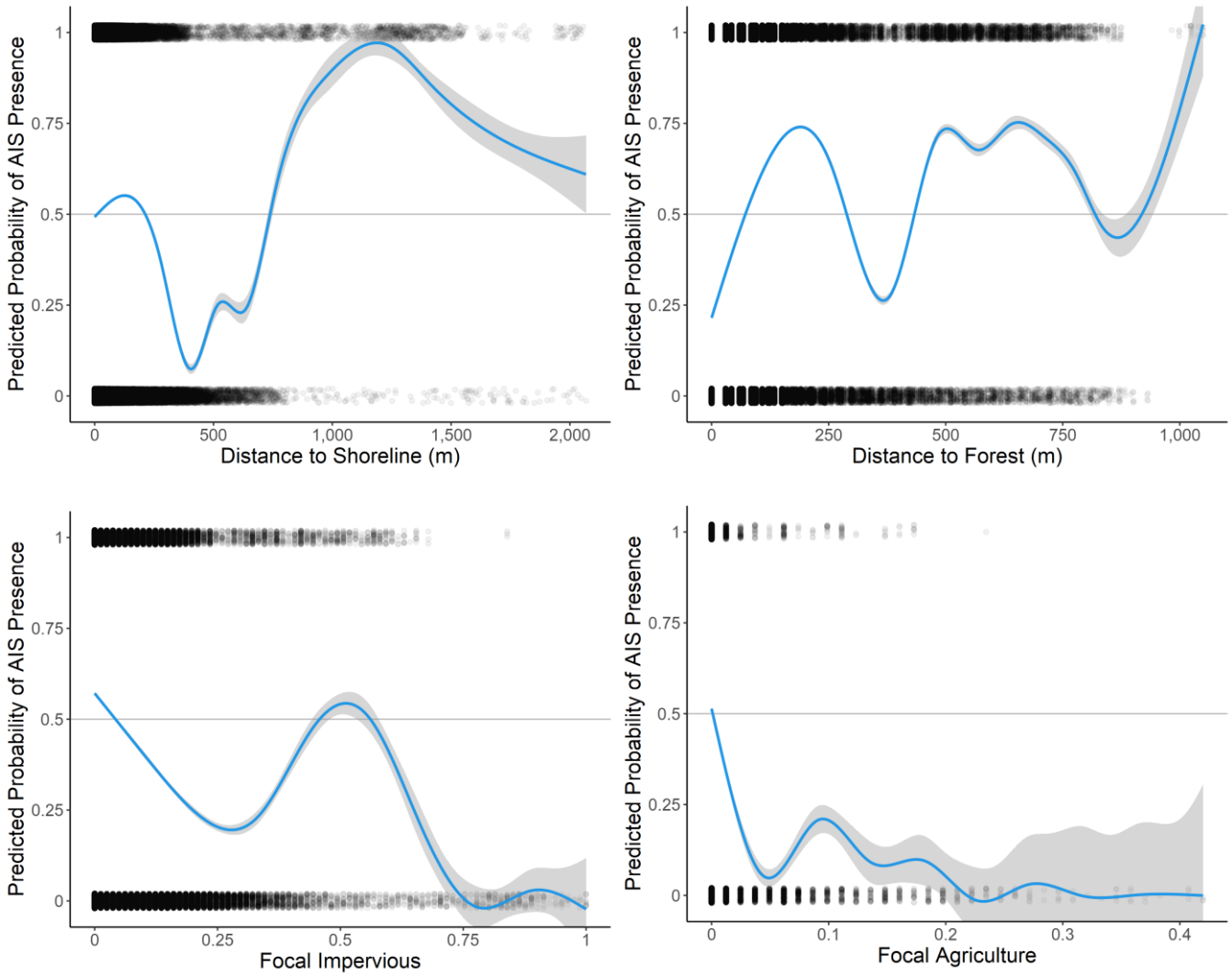


Figure 10. Partial dependence plots for the four predictor variables. The blue line indicates the predicted probability of AIS at a given value for the predictor variable of interest, and the black points represent data in the training dataset (points are jittered vertically to enable visualization of multiple overlapping points).

While the model performance of the linear model (LASSO) was worse than that of the XGBoost best subsets models, interpretation of the coefficients of linear models is straightforward and thus may lend insights despite the lackluster model performance. LASSO output indicated that the points of interest variables had the most predictive power of AIS (Figure 11).Variables that were strongly positively associated with AIS included distance to hand boat launch (group 7), distance to beach (group 9), distance to marina (group 6), and distance to agricultural land cover. The interpretation from this outcome is the closer a given location is to these areas, the lower the AIS probability. Variables that were strongly negatively correlated with AIS included distance to trailer boat launch (group 8), distance to fishing location (group 5), distance to impervious anthropogenic surface (group 1). The interpretation from this outcome is the closer a given location is to these areas, the higher the AIS probability.

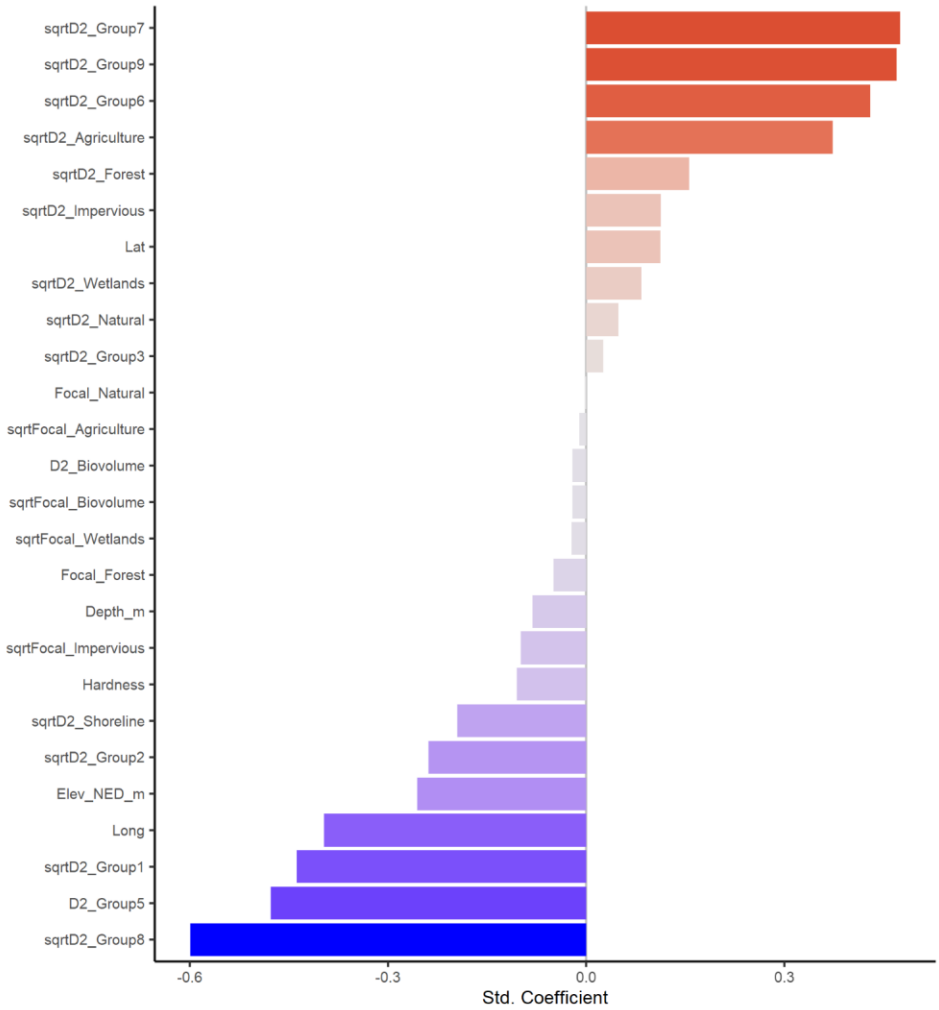
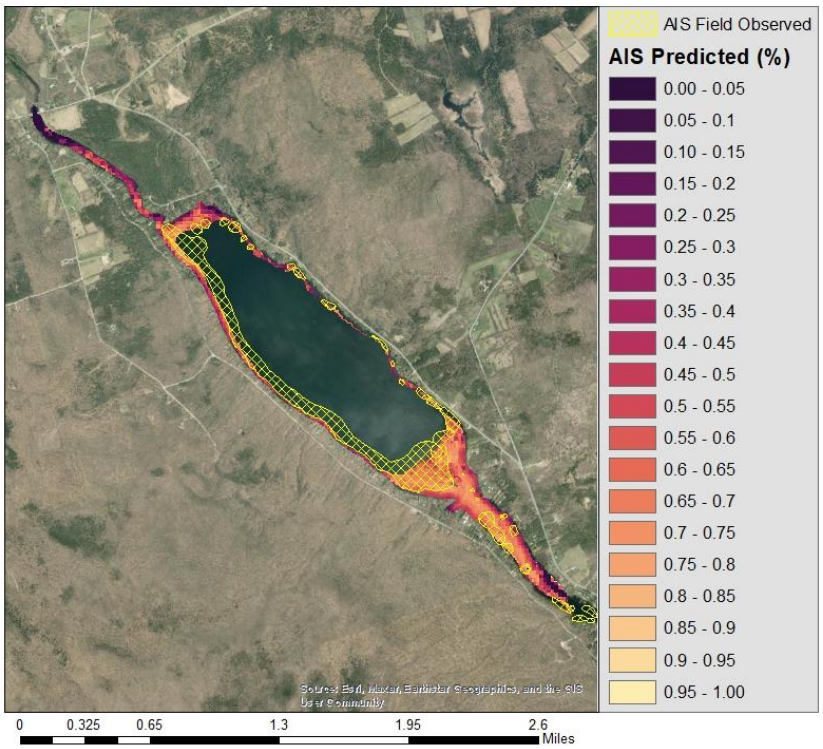


Figure 11. Predictor variable standardized coefficients from the LASSO model.

3.4 SELECTION OF BEST-PERFORMING MODEL

The XGBoost “best subset” models had the best model performance. The top-performing model was one that included distance and focal biovolume metrics, but this model only marginally outperformed the best subset model that excluded biovolume. The slight improvement in model performance may not be worth the cost of obtaining biovolume data in terms of labor, variable data quality, and expense of service. Additionally, the XGBoost model without biovolume was slightly better at predicting positive AIS presence and had fewer false negatives, both key factors in how this model would be used to aid field surveys. **Thus, the XGBoost best subset model without biovolume was chosen as the best-performing model.**

In addition to the metrics evaluated at the individual pixel level (Table 4), the adjacency of correctly and incorrectly predicted pixels to areas of confirmed AIS was also an important indicator of model performance. Pixels located within areas of confirmed AIS presence tended to have the highest predicted probabilities, and pixels with the lowest predicted probabilities were located far away from areas of confirmed AIS presence. Importantly, false positives tended to be located near areas of confirmed AIS cover and could represent areas with suitable habitat that are vulnerable to future invasion (Figure 12).



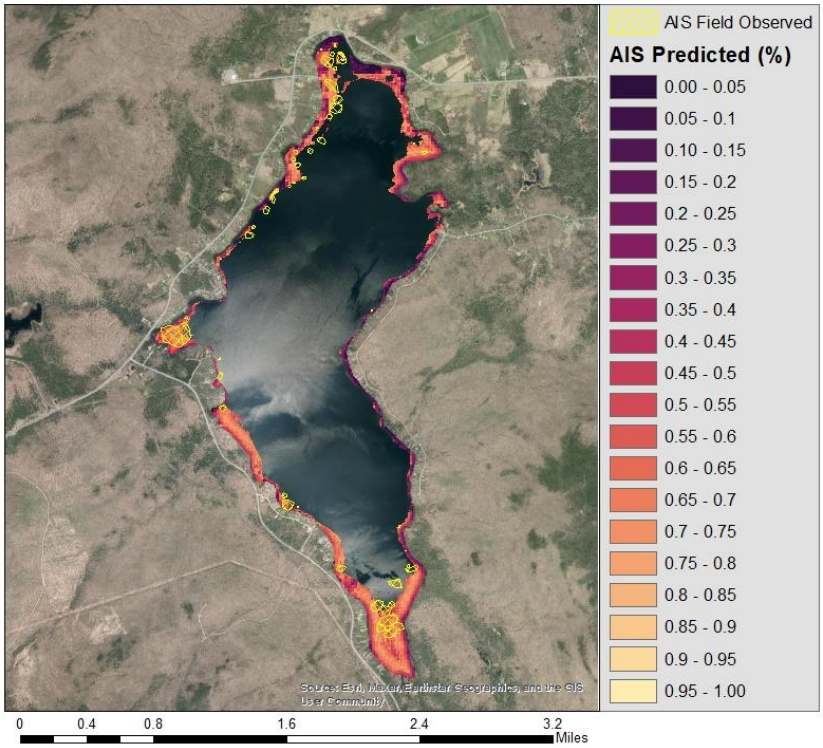


Figure 12. Example model output for Lower Chateaugay Lake (top) and Chazy Lake (bottom), two lakes in the model training dataset. Areas with confirmed AIS presence are noted in yellow hatching, and predicted AIS probabilities are shown with the color scale.

3.5 MODEL APPLICATION TO LAKES IN ADIRONDACK PRISM

Lakes in the Adirondack PRISM boundary were associated in a similar geographic context as the testing and training lakes, and comparisons of the distributions for the four predictor variables also indicated similar conditions (Figure 13). One potential limitation was the extrapolation of the model to lakes whose areas are larger than the training dataset (i.e., Lake George, Great Sacandaga Lake, Cranberry Lake), so the results for these lakes should be interpreted with caution. The predicted probability of AIS cover in the lakes across the Adirondack PRISM ranged from 0-100%, with a similar interquartile range and median prediction as the training dataset (Figure 14). Lakes in the testing dataset tended to have narrower distributions than the training and application groups across most variables.

One caveat of the model application is that the model was trained on a subset of lake pixels that were <7 m deep, which necessitated detailed bathymetry input data. Although depth was not chosen by the model as an important predictor, the lack of bathymetry data for the large majority of lakes in the Adirondack PRISM meant that the 7 m cutoff was unable to be applied. Therefore, AIS predicted probabilities were generated for entire lake areas and should be interpreted with caution and ideally alongside field-deployed depth finders if being used for monitoring purposes (see Figure 15 for examples). An interactive map was developed to show model results.

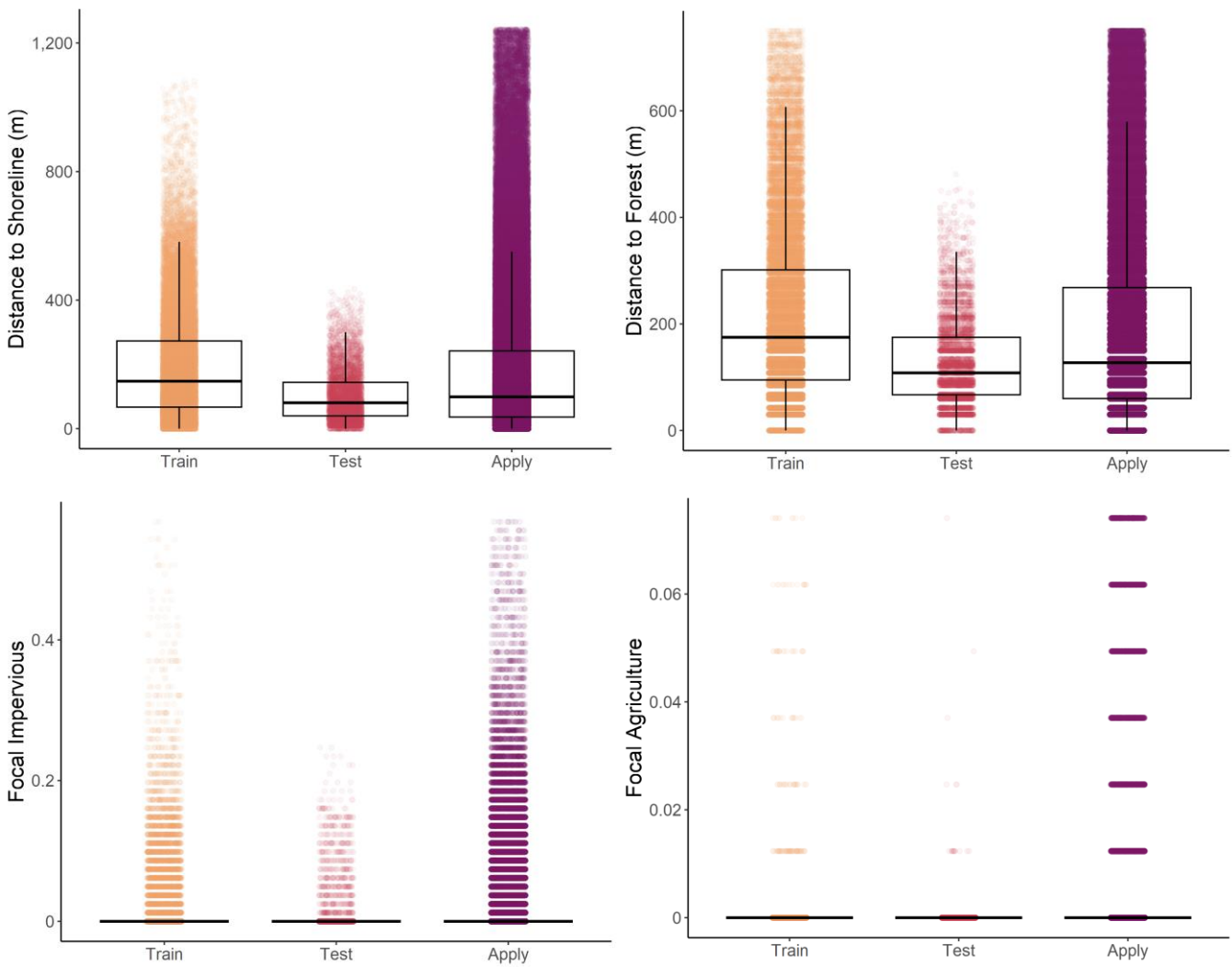


Figure 13. Distributions of the four predictor variables in the XGBoost best subset model for the three lake groups.

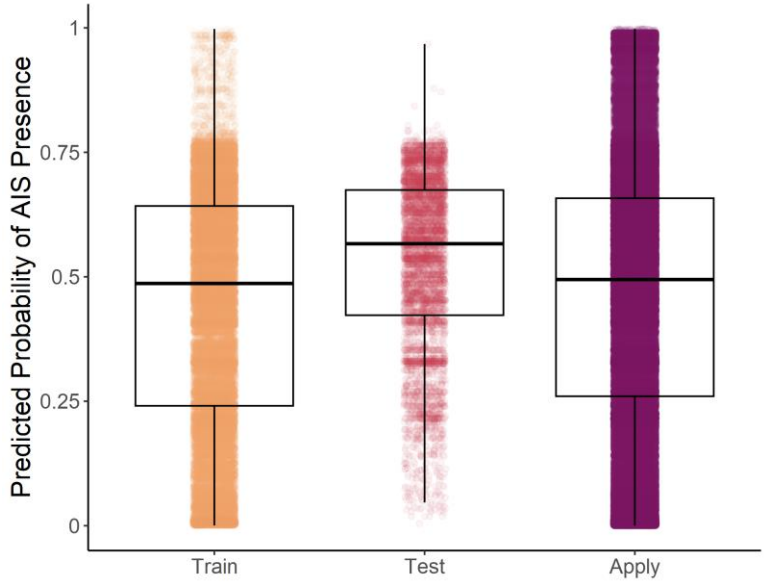


Figure 14. Predicted AIS probabilities across the three lake groups.

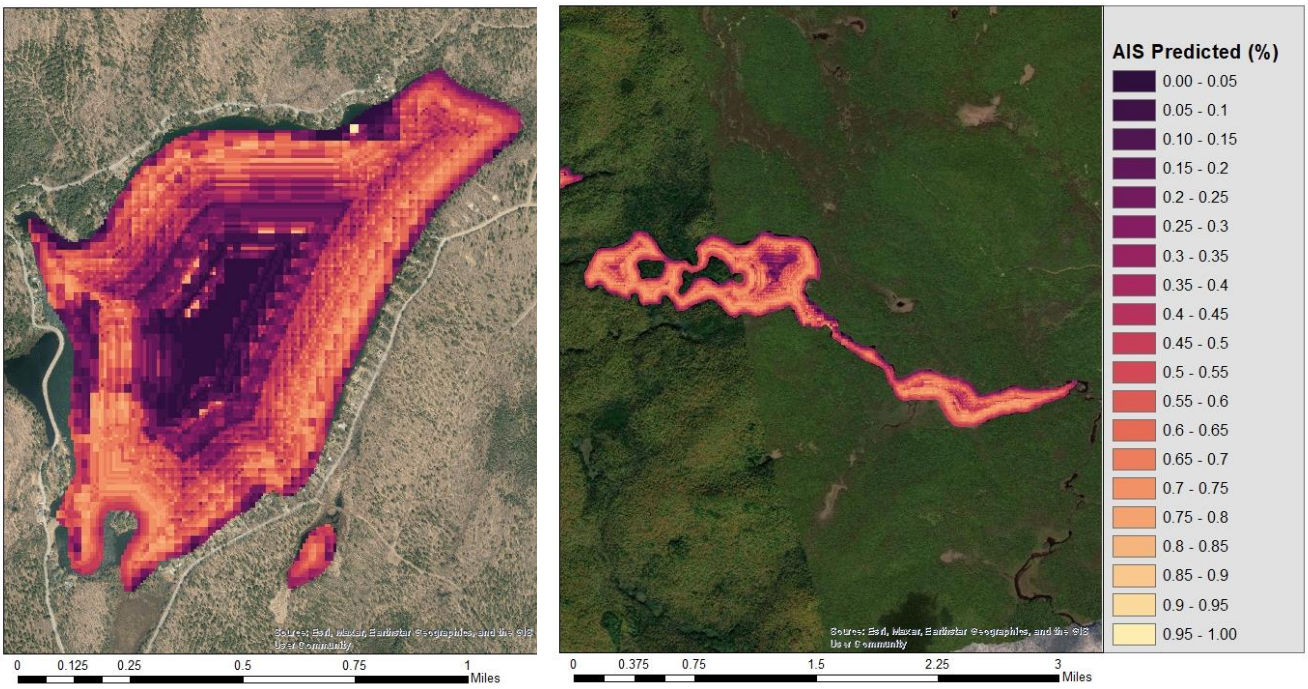


Figure 15. Example model output for Fern Lake (left) and Newcomb Lake (right), two lakes in the model application dataset. Predicted AIS probabilities are shown with the color scale. Note that in the absence of bathymetry data, no areas of the lake were left out of the model output due to depth.

4.0 CONCLUSION

A modeling framework was developed to evaluate the variables and model structure that best predicted AIS presence within lakes in the Adirondack PRISM. To prepare for model development and calibration, a comprehensive database was compiled that included within-lake data (i.e., water depth, hardness, biovolume, and distance to shoreline) as well as adjacent land cover and anthropogenic points of interest data that may be relevant to AIS cover and invasion. The core lakes within this database were those that have been monitored for AIS cover and have also been mapped with sonar-derived variables using BioBase. The database also included data from lakes across the Adirondack PRISM. Multiple models were tested, including a linear model, a tree-based machine learning model, and an artificial neural network. Model performance was tested head-to-head, which enabled the identification of the best model structure to predict within-lake AIS cover. The top two best-performing models were tree-based machine learning model (XGBoost) which incorporated factors such as biovolume, distance to shoreline, distance to forest cover, density of adjacent impervious cover, and density of adjacent agricultural cover as predictor variables. While it was initially surprising that some variables were not important predictors of AIS (e.g., proximity of boat launches, water depth), it was hypothesized that those variables were collinear to some degree with the variables of highest importance. For instance, distance to shoreline likely accounts for some of the variability in water depth. The results of the within-lake modeling effort were consistent with previous efforts to predict AIS vulnerability across lakes, particularly the findings that development was associated with higher vulnerability and forested land cover was associated with lower vulnerability (Shaker et al. 2017). Ultimately, the XGBoost model that did not include biovolume variables as predictors was selected as the top model and was applied to additional lakes that had unknown AIS presence. This model was selected to build an interactive webmap of predicted AIS cover across lakes in the Adirondack PRISM based on performance criteria and ability to predict vulnerability across thousands of lakes in the region using national-scale spatial datasets.

Evaluation of the individual model predictors on their own enabled an actionable interpretation of the importance of these predictors for AIS monitoring purposes. AIS presence was predicted in areas within 200 m of the shoreline and was more likely in areas with a high density of impervious and/or agricultural cover. Distance to forest was also an important predictor, with a bimodal distribution with highest probabilities predicted at ~200 and 500-750 m from forest cover. Alongside these general predictions, the geospatial layer displaying predicted AIS probability across the >1000 lakes in the Adirondack PRISM enables location-specific predictions within lakes. While this model does not replace the need for physical monitoring of AIS, the data-enabled pairing of machine learning model predictions with on-the-ground monitoring efforts will allow for more efficient identification of areas likely to have AIS plant cover and potential areas vulnerable to future AIS plant invasion.

5.0 REFERENCES

Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. DOI: 10.1145/2939672.2939785.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

Friedman, J.H., Hastie, T., and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL https://www.jstatsoft.org/v33/i01/.

R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna,Austria. URL https://www.R-project.org/.

Shaker, R.R., Yakubov, A.D., Nick, S.M., Vennie-Vollrath, E., Ehlinger, T.J., and Forsythe, K. W. 2017. Predicting aquatic invasion in Adirondack lakes: a spatial analysis of lake and landscape characteristics. Ecosphere 8(3):e01723. DOI: 10.1002/ecs2.1723

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

APPENDIX A: POINTS OF INTEREST METADATA

Group 1 - Impervious Anthropogenic Sources	Source
Bakery	ADK Atlas
Cabins and Cottages	ADK Atlas
Club	ADK Atlas
Cross-Country Ski Center	ADK Atlas
Deli	ADK Atlas
Gas	ADK Atlas
Golf	ADK Atlas
Hostel	ADK Atlas
Hotel	ADK Atlas
Ice Rink	ADK Atlas
POI	ADK Atlas
Restaurant	ADK Atlas
Retail Store	ADK Atlas
Snack Bar	ADK Atlas
Summer Camp	ADK Atlas
Tavern	ADK Atlas
Theater	ADK Atlas
Vacation Rentals	ADK Atlas
Visitor Center	DEC
Observation Platform	DEC
Observation Tower	DEC
Fire Tower	DEC
Scenic Vista	DEC
Group 2 - Camping	**Source**
Campground boundaries/pts	DEC
Day use area	DEC
Picnic area	DEC
Picnic table	DEC
Lean to	DEC
Campground	ADK Atlas
Leanto	ADK Atlas
Group 3 - Pervious Anthropogenic Sources	**Source**
Ball Field	ADK Atlas
Equestrian	ADK Atlas
Public Park	ADK Atlas
Scenic Overlook	ADK Atlas
Trail Access	ADK Atlas
Equestrian Platform	DEC

Group 4 - Parking Lots	Source
Paved Parking Lot	DEC
Unpaved Parking Lot	DEC
Group 5 - Fishing	**Source**
Fishing Access Site	DEC
Fishing Pier	DEC
Fishing Platform	DEC
Fishing Access Site	ADK Atlas
Water Access	ADK Atlas
Group 6 - Marina	**Source**
Marina	ADK Atlas
Group 7 - Boat Launch (Hand)	**Source**
Hand Launch	DEC
Hand Launch	ADK Atlas
Group 8 - Boat Launch (Trailer)	**Source**
Ramp Launch	DEC
Trailer unimproved boat launch	ADK Atlas
Trailer Improved Boat Launch	ADK Atlas