Prioritization of aquatic invasive species surveillance sites for inland lakes in Great Lakes states

Regional Aquatic Invasive Species Prevention, Early Detection and Response
Phase IV (2020 – 2022)

Summary

This report summarizes the activities and outcomes related to Objective 3 of the Great Lakes region's Aquatic Invasive Species Early Detection and Response – Phase IV project, agreement no. F21AP00368. With technical assistance from others in the Great Lakes states doing similar work, we developed a model for prioritizing AIS monitoring sites across inland lakes. We collected and synthesized observations of AIS introduced to the region relatively recently and developed region-wide spatial data sets to serve as surrogates for major pathways of AIS introduction and habitat suitability. We then developed generalized, machine-learning-based models for invasive aquatic plants, invertebrates and fish to identify potential hotspots for early introductions across approx. 68,000 inland lakes in the Great Lakes states. The results are available for browsing and download via an interactive web app.

1. Introduction

In recent decades, an increasing number of freshwater species have come to be considered invasive, and some have caused widespread environmental effects and economic harm. The Great Lakes region of North America is particularly vulnerable to the introduction and associated impacts of aquatic invasive species (AIS), with 194 nonindigenous aquatic species established or reported within the Great Lakes basin (NOAA 2024). A 2016 study estimated the aggregate cost of AIS to the Great Lakes region at over $100 million annually (Rosaen et al. 2016).

In the Great Lakes States and Tribal territories, AIS surveillance programs are primarily directed towards monitoring inland lakes and river systems. The early detection of and rapid response to AIS invasions of inland waterways is also a major determinant of AIS invasions into the Great Lakes themselves. To be effective, it is important for these programs to optimize the likelihood of detecting AIS occurrences while they are still at a stage where they can be extirpated or controlled (Keller et al. 2008, Vander Zanden et al. 2010). One of the primary challenges when developing early detection and rapid response (EDRR) programs is how to prioritize monitoring sites and develop sampling strategies while maximizing the use of fixed budgets and personnel capacity. Here, we aimed to develop a

generalized, machine-learning-based modeling approach to identify hotspots for early introductions of AIS within inland lakes in the Great Lakes states (Illinois, Indiana, Michigan, Minnesota, New York, Ohio, Pennsylvania, and Wisconsin).

1.1 Technical workshop

To provide direction for the development of our inland lakes surveillance prioritization model, a technical workshop was convened virtually on February 9, 2023 to review current systems in use across states and Tribes to prioritize inland AIS management. Attendees were also asked to provide technical input into the approach for a regional model that would complement the work already being done within state and regional/national programs.

Table 1 lists the workshop attendees.

Table 1: Attendees of February 2023 technical workshop on prioritization modeling

ATTENDEE	AFFILIATION
LINDSAY CHADDERTON	The Nature Conservancy
CECI WEIBERT	Great Lakes Commission
AMANDA GRIMM	Great Lakes Commission
JENNIFER DEAN	New York Department of Environmental Conservation
NICHOLAS PHELPS	University of Minnesota
AMY KINSLEY	University of Minnesota
PAUL VENTURELLI	Ball State University
JONATHAN BOSSENBROEK	University of Toledo

Tools for prioritization of AIS monitoring across inland lakes have already been created at the state level for two states in the Great Lakes region. Jennifer Dean presented to the group on the prioritization tool developed by the New York Natural Heritage Program (NYNHP), while Nicholas Phelps and Amy Kinsley presented on the University of Minnesota's AIS Explorer. The NYNHP tool allows users to select their own risk thresholds for AIS introduction, establishment, and impact to identify the subset of lakes most important for monitoring based on their own priorities (NYNHP 2022). Introduction risk is characterized by variables such as modeled boat traffic and proximity to known infestations, establishment risk by species distribution models (SDMs) for eight individual AIS selected to represent major functional groups, and impact risk using several metrics of ecological condition, recreational value, and stressors. This approach was directed by a steering committee representing several state, federal, and regional agencies, but it focuses on priorities specific to New York and relies on datasets that are not available

across the other Great Lakes states, such as The Nature Conservancy's Northeast Lake and Pond Data (Olivero-Sheldon et al. 2014).

The other existing state-level tool, the University of Minnesota's AIS Explorer, estimates and scores introduction risk for individual lakes and ponds in Minnesota for two indicator species: zebra mussel and starry stonewort (Kinsley et al. 2022). The scores in this tool are based on a combination of estimated habitat suitability, hydrological connectivity with other waterbodies, and a network-based model of boater movement. With its focus on boats as a dispersal vector, this model focuses on introduction risk and does not include evaluation of establishment risk or invasion impact. An expansion of this model to Wisconsin is currently underway. Major characteristics of both state-level models are summarized in Table 2. Details were also shared by Paul Venturelli on the work they had underway exploring whether data from fishing apps can be used to provide a standard regional measure of recreational use of, and connectivity between, individual lakes across the Great Lakes states. This information was used to refine the approach and objectives for this modeling effort.

SDMs have been used by a number of groups as a tool for identifying areas at greater risk of invasion and planning surveillance strategies (e.g., Bazzichetto et al. 2018, Lozano et al. 2023). These models, typically developed for an individual species, predict the species' potential range based on the environmental characteristics of known occurrence locations. A downside of this approach is the time and expertise required to obtain the needed species-specific data and calibrate/validate each model. Due to these requirements, SDM-based analyses are only available for selected, priority invasive species. A generalized approach that could help to identify areas that are at higher overall risk of AIS introductions, especially early in a species' spread into a region, without requiring species-specific model development could greatly improve data-driven surveillance and management.

We used AIS presence records from multiple observation databases combined with potential predictor variables representing aspects of habitat suitability and invasion pressure to determine inland lakes that might be prone to introductions of AIS to help predict where future inland invasions in this region would be most likely to occur, possibly also for species not included in our data. An extreme gradient boosting (XGBoost) model was trained to predict new introductions of AIS as a function of the predictor variables for each of three broad guilds: aquatic plants, invertebrates, and fish. Our analysis shows that this approach can produce hotspot maps for likelihood of AIS detection at a regional scale and with a high predictive accuracy, but with important limitations.

Table 2: Summary of the existing AIS surveillance prioritization tools for lakes in the Great Lakes states.

MODELING TOOL	AQUATIC INVASIVE SPECIES POND AND LAKE VULNERABILITY PRIORITIZATION TOOL FOR NEW YORK	MINNESOTA AIS EXPLORER
URL	https://www.nynhp.org/projects/aquatic-invasive-prioritization/	https://www.aisexplorer.umn.edu/#!/map
INDICATOR SPECIES FOR MODEL	Indicator species selection was guild-based: Starry stonewort (submerged algae) Hydrilla (submerged plant) EWM (submerged plant) Water chestnut (floating plant) Northern snakehead (fish) Rusty crayfish (benthic invertebrate) Spiny waterflea (planktonic invertebrate) Zebra mussel (bivalve)	Starry stonewort EWM Spiny waterflea Zebra mussel
MODEL SUMMARY	Threshold-based assessment framework building on Vander Zanden and Olden (2008). Risk score, suitability level, and invasion impact score are calculated in advance for each lake + species. User can select thresholds for all three factors and output a list of 'lakes at risk of establishment'. If multiple species are selected, the resulting 'lakes at risk' include all lakes that meet the criteria for any selected species.	Network-based optimization algorithm that selects watercraft inspection sites to maximize the number of inspections of watercraft that move from AIS infested to uninfested lakes. Output for a selected indicator species is based on known infested lakes + boating data. If multiple species are selected by the user, they are weighted equally.
PROXIES OF INVASION PRESSURE	Proxy variables were selected to represent introduction risk via the following: • boats (modeled boat traffic using WISPA steward data for one high-use weekend in 2019 to predict relative traffic for all lakes/ponds in dataset) • people (average annual daily traffic on roads near a pond/lake based on DOT data; not every road included) • dispersal from existing AIS locations (straight line and flowline distances to nearest known location) dispersal via	UMN's state-level database of estimated boat traffic between lakes; MN DNR data on lake invasion status

MODELING TOOL	AQUATIC INVASIVE SPECIES POND AND LAKE VULNERABILITY PRIORITIZATION TOOL FOR NEW YORK	MINNESOTA AIS EXPLORER
	waterfowl (number of unique AIS within a search radius based on waterfowl foraging distances).	
PROXIES OF ESTABLISHMENT RISK	SDMs for selected species representing ecological guilds. Each lake assigned low, medium or high suitability for each species	Not incorporated
PROXIES OF POTENTIAL IMPACT	• Water/habitat quality: Landscape condition assessment, biological assessment profile, native fish richness, HABs • Rare species/communities: natural heritage program occurrences	Not incorporated
WEB DASHBOARD PLATFORM	Shiny	Shiny
SELECTOR WIDGETS AVAILABLE IN WEB DASHBOARD	Area filter	Species
	Introduction risk threshold	Which risk score (boater movement, water connectivity, or combined)
	Selected species and suitability thresholds for each	Risk score slider to threshold results
	Impact score threshold	Filter to one county/tribal land
USEFUL DASHBOARD FUNCTIONS	Download results as zipfile with csv and gpkg format options	Algorithm responsive to user input, outputs downloadable as image or table, separate map layer of lakes infected with the selected species, updates to MN DNR infestation data automatically pulled into dashboard

MODELING TOOL	AQUATIC INVASIVE SPECIES POND AND LAKE VULNERABILITY PRIORITIZATION TOOL FOR NEW YORK	MINNESOTA AIS EXPLORER
	Can generate a report in html format that documents the criteria that the user selected and how many lakes qualified using those criteria	A second tool (Prioritization for watercraft inspections) generates a list of the inspection stations needed to inspect a user-defined % of risky boats in the area

2. Methods

2.1 Data

2.1 Spatial framework

The model was limited to inland lakes within the Great Lakes states. 'Lakes' included both natural lakes and reservoirs with defined polygons in the moderate-resolution National Hydrography Dataset Plus version 2 (NHDPlus v2) and an area > 3 ha (approx. 7.4 acres), a total of 46,764 lakes. NHDPlus v2 was used rather than NHDPlus High Resolution because our environmental variables included several metrics that had previously been calculated for NHDPlus v2 lakes by other researchers and that we did not have the capacity to develop for NHDPlus HR lakes. The minimum lake size was extended from our initially planned 10 acres because the data for the smaller lakes was readily available. Lakes for which complete environmental data was not available were excluded, resulting in 24,813 modeled lakes.

Species observation and boat launch locations were snapped to the NHDPlusV2, and related connectivity metrics were calculated for each modeled lake using the PyNHD python package to access several USGS web services, trace NHD flowlines upstream from each lake, and calculate stream distances. Species observations with locations that do not fall within an NHD waterbody polygon were assigned to the nearest NHD waterbody or flowline within 500 m. Observations more than 500 m from any NHD feature were not excluded, as the dataset spans several decades and hydrology has changed over time, but they were not included in the calculation of connectivity metrics that use species observation locations.

2.1 Species presence

A list was assembled of aquatic plant, invertebrate, and fish species that are regulated at the state or federal level in the Great Lakes region and/or are included in the Great Lakes St. Lawrence Governors and Premiers' "Least Wanted" AIS list (GSGP 2019). Species were then dropped from this list if they have not yet been observed within the region, if they are range expanders that are native to a portion of the region (in order to develop a more consistent region-wide model), or if the species was already well-established in the region prior to 1970. This last criterion was applied to keep the focus on current-day invasion risk by invasive species novel to the region. Earlier invaders (such as curly-leaf pondweed, probably introduced to the Great Lakes region around the 1880s) have had decades to expand into lower-risk waterbodies and likely experienced a very different pattern of spread prior to the expansion of the interstate highway system, an important long-distance

dispersal corridor (Weir et al. 2022). As a result, observations of their current distribution may be less predictive of future early detection hotspots.

Observations of the species of interest were acquired from five databases that compile species observation data: MISIN, GLANSIS, iMapInvasives, GBIF, and EDDMapS. For the results shown in this report, all observation datasets were accessed on Jan 17, 2025. The raw data from all sources was cleaned by removing all points with any of the following characteristics: missing observation date, reporting year prior to 1970, location outside of the Great Lakes states, or location falling exactly on biodiversity institutions, assuming that those are part of a collection (Zizka et al. 2019). Furthermore, presences with duplicate values for species name, observation date, and location (with a small buffer built in to handle possible differences in coordinate rounding across databases) were flagged and apparent duplicates were dropped. Proximity to surface water was not used as a QA/QC measure because the observations span several decades, and we don't have consistently reliable maps of surface water extent dating back to the beginning of the dataset. The final list of included species and the numbers of observations available after QA/QC are provided in Table 3. To the extent possible, additional QA/QC measures were applied to the observations from each database based on its attributes with the goal of restricting the final dataset to research-grade observations.

2.1.1 MISIN

The Midwest Invasive Species Information Network (MISIN) is a regional data aggregation effort led by researchers at Michigan State University to develop and provide an early detection and response resource for invasive species in the Midwest region (Higman et al. 2010). Observations are accepted from both experts and citizen scientists. MISIN data was collected from a REST service that is maintained by MISIN and updated daily. The data were further cleaned by keeping only the records flagged as verified by a trusted source.

2.1.2 GLANSIS

GLANSIS is a NOAA-led, inter-agency database providing focused information on non-native aquatic species found in the Great Lakes region (Sturtevant et al. 2022). NOAA-GLANSIS staff collect and curate verified reports within the US Great Lakes basin for all species on the GLANSIS nonindigenous list, which is limited to aquatic species that are present in the Great Lakes basin below the ordinary high-water mark. GLANSIS data were accessed as a zipped CSV file on the GLANSIS website updated weekly. The data was further cleaned by dropping observations for which the coordinates were flagged as a centroid rather than a more precise location.

2.1.3 iMapInvasives

The iMapInvasives platform is developed by NatureServe and used by a variety of users including natural resource agencies in selected states to share invasive species data. Within the Great Lakes, New York and Pennsylvania are participating jurisdictions. Data were obtained through iMapInvasives' REST services. Observations available through the mapping services have been confirmed by either the state administrator or a project leader. The dataset does not include any information that lends itself to further cleaning.

2.1.4 GBIF

The Global Biodiversity Information Facility (GBIF; Edwards 2004) is a worldwide open-access database of species distribution data, not limited to invasive species. It aggregates occurrence records from more than 1100 institutions around the world and publishes these in a standardized format. The GBIF dataset was obtained through their API using the pygbif package (Chamberlain and Boettiger 2017) and further cleaned by removing observations based on a fossil specimen, preserved specimen, and literature-based observations as well as those flagged by GBIF as having geospatial issues.

2.1.5 EDDMapS

EDDMapS is another web-based invasive species mapping system based at the University of Georgia (EDDMapS 2025). All data submitted to EDDMapS is reviewed for accuracy by state verifiers. The majority of observations in the EDDMapS dataset for our species of interest do not have the exact coordinates shared publicly and so were not usable here. EDDMapS observations were retained if the "identification credibility" value for the observation was either "Credible" or "Verified".

Table 3: Observations of focal species within Great Lakes states used for modeling.

TAXON	SCIENTIFIC NAME	COMMON NAME	OBSERVATIONS
FISH	Channidae	Snakeheads	254
	Clarias batrachus	Walking catfish	0
	Gymnocephalus cernua	Ruffe	822
	Misgurnus anguillicaudatus	Pond loach	122
	Neogobius melanostomus	Round goby	3447
	Osmerus mordax	Rainbow smelt	561
	Osteoglossum bicirrhosum	Silver arowana	0
	Proterorhinus semilunaris	Tubenose goby	119
	Tinca tinca	Tench	12
	Total		**5337**
PLANTS	*Alternanthera philoxeroides*	Alligator weed	1
	Cabomba caroliniana	Carolina fanwort	2252
	Callitriche stagnalis	Pond water-starwort	42
	Egeria densa	Brazilian waterweed	487

	Hottonia palustris	Water violet	2
	Hydrilla verticillata	Hydrilla	1857
	Hydrocharis morsus-ranae	European frog-bit	11738
	Hygrophila polysperma	Indian swampweed	8
	Limnophila sessiliflora	Dwarf ambulia	0
	Ludwigia grandiflora, L. hexapetala, L. peploides	Large-flower, six petal, and creeping water primrose	1, 1, 387
	Marsilea mutica	Australian water clover	1
	M. quadrifolia	European water clover	260
	Myriophyllum aquaticum	Parrot feather	280
	Najas minor	Brittle naiad	2666
	Nasturtium officinale	Water-cress	1113
	Nelumbo nucifera	Sacred lotus	66
	Nitellopsis obtusa	Starry stonewort	8403
	Ottelia alismoides	Duck-lettuce	0
	Pistia stratiotes	Water lettuce	458
	Pontederia azurea, P. crassipes	Anchored and common water-hyacinth	0, 193
	Sagittaria sagittifolia	Arrowhead	0
	Salvinia auriculata, S. molesta	Eared and giant salvinia	0
	Spirodela punctata	Dotted duckweed	1
	Stratiotes aloides	Water soldier	1
	Trapa natans	European water chestnut	11846
	Total		**42,064**
INVERTEBRATES	*Bithynia tentaculata*	Faucet snail	541
	Bythotrephes longimanus	Spiny water flea	172
	Cercopagis pengoi	Fishhook water flea	32
	Corbicula fluminea	Basket clam	4081
	Dreissena bugensis	Quagga mussel	970
	Dreissena polymorpha	Zebra mussel	4852
	Eriocheir sinensis	Mitten crab	106
	Hemimysis anomala	Bloody red shrimp	78
	Melanoides tuberculata	Red-rimmed melania	3
	Potamopyrgus antipodarum	New Zealand mud snail	265
	Procambarus virginalis	Marbled crayfish (Marmokrebs)	0
	Total		**11,100**

We flagged observations that were present in more than one of these databases, which allowed us to evaluate both the volume of observations available through each database and the extent to which the databases overlap. Figure 1 shows the total number of usable observations available from each source as well as the proportion of that total that did not appear in any of the other databases (unique observations). iMapInvasives contributed the largest number of observations, including a large number of European water chestnut

observations in New York, even though the iMapInvasives data was mostly limited to New York and Pennsylvania. The available observation data was also strongly skewed towards a handful of the species of interest (Figure 2).

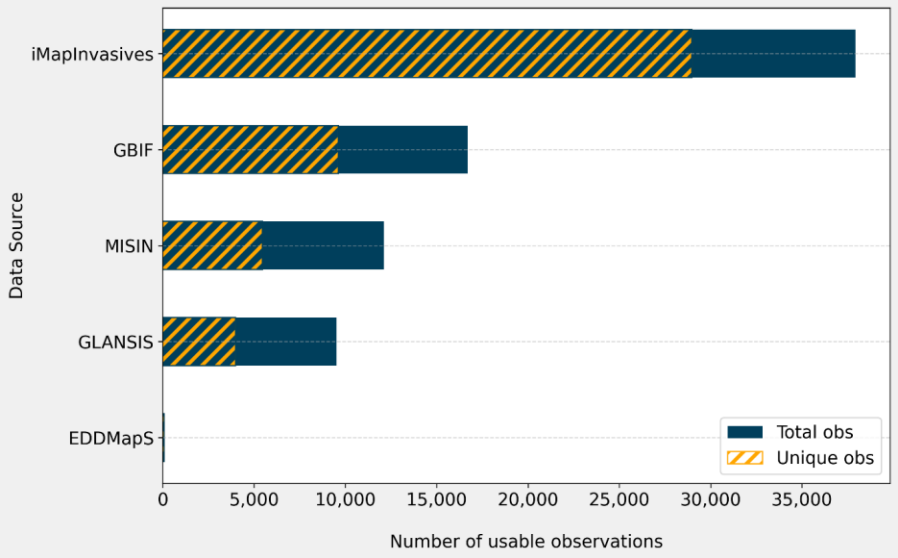

Figure 1: Numbers of species observations available for the model from different sources.

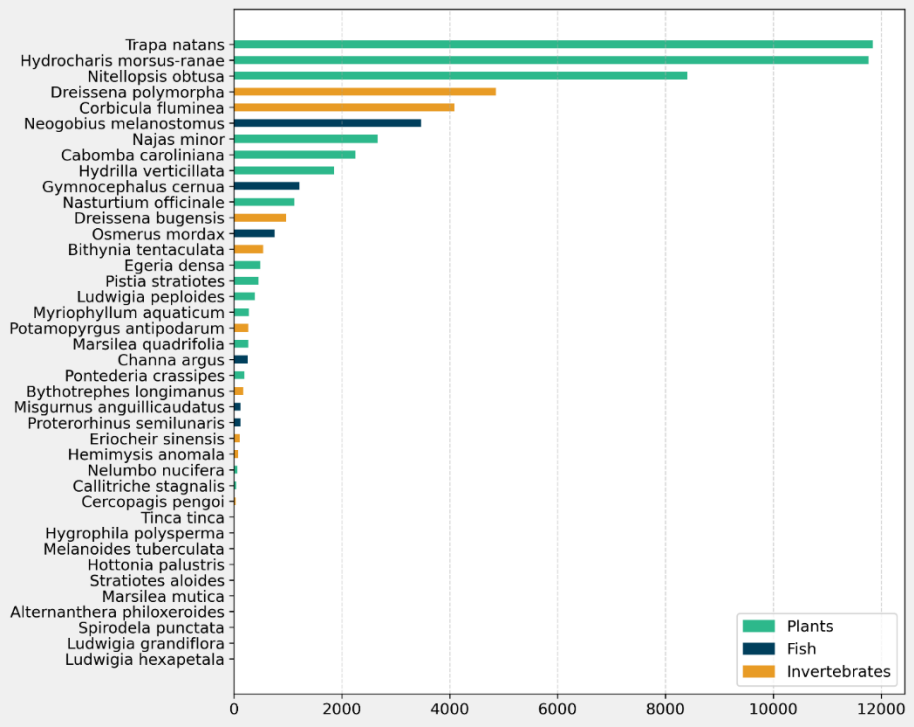

Figure 2: Number of observations collected per species. Bars are colored by major taxon, representing the three generalized models that were developed.

2.1.2 Background data

Some of the species observation databases from which we acquired presence data also collect reports of species absences, but these absence observations are limited and mostly do not include information about sampling effort. Observations (and presumably sampling effort) were also spatially clustered near urban areas, especially the metro Detroit and Finger Lakes areas for plant observations, the Finger Lakes area for invertebrate observations, and the Duluth area for fish observations (Figure 3). To help address these issues and to ensure methodological consistency, we generated pseudo-absences for our background data using an observer-oriented approach (Milanesi et al. 2020). A list of observers of the focal species was generated from the presence data. Assuming that a given observer of a given species would collect locations of such species when they encounter it in the field, locations with reported occurrences of species other than the focal species collected by those same observers were considered as pseudo-absences. Milanesi et al. found that this approach improved species distribution models relative to random pseudo-absences.

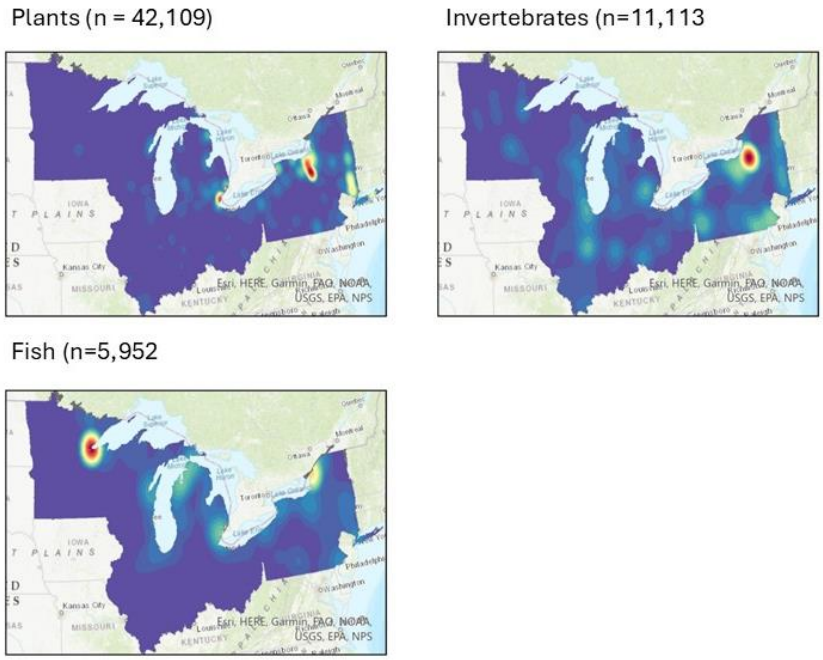

Figure 3: Density of focal species observations, calculated using a kernel function.

2.1.3 Risk variables

We collected or developed 44 predictor variables that were contiguously available for the entire region. Variables were selected to serve as surrogates for invasion pressure and direct/indirect habitat suitability for AIS.

- Model inputs to represent invasion pressure included metrics related to boat access locations, modeled freshwater fishing demand, nearby reported presence locations, distance to the nearest highway, surrounding population density, hydrological connectivity, and recreational boating connectivity.
- Habitat suitability was predicted based on metrics related to lake size, 30 year average temperature range, water quality, whether the lake is an artificial reservoir, presence of known dams, surrounding land cover, and disturbance level,
- Additional variables included lake area and shape, latitude, distance from the Great Lakes, total number and area of connected lakes, catchment and watershed area, and the stream and canal/ditch density of the watershed.

Table 4 summarizes these variables and their sources.

VARIABLE	DESCRIPTION	SOURCE
AGENCY_ALL_PCT	Percent of watershed that is protected areas from PAD-US	LAGOS-US GEO (Smith et al. 2022)
ANGLER_INC	Weighted incoming connections to a lake according to angler network model, from https://bsumaps.maps.arcgis.com/home/item.html?id=9851db4d6d1c4e7b9426d37972069eb2	calculated by BSU for this project
AREA	Lake surface area in ha	LAGOS-US GEO (Smith et al. 2022)
BOAT_PARKING	boat ramp parking lot size (number of spaces); 0 = no boat ramp	EPA Region V Inland Sensitivity Atlas; PA Fish and Boat Commission Access Points (PAFBC 2022); NYSDEC Boat Launches (NYSDEC 2022)
CANALS_MPERHA	canal and ditch density defined as the sum of all canals and ditches within the zone divided by the area of the zone	LAGOS-US GEO (Smith et al. 2022)
CLIMATE8110NORM_TMAX_DEGC	mean value within the zone or value at lake central point for the 30 year normal annual maximum temperature; calculated for data from 1981 to 2010	LAGOS-US GEO (Smith et al. 2022)
CLIMATE8110NORM_TMEAN_DEGC	mean value within the zone or value at lake central point for the 30 year normal annual mean temperature; calculated for data from 1981 to 2010	LAGOS-US GEO (Smith et al. 2022)
CLIMATE8110NORM_TMIN_DEGC	mean value within the zone or value at lake central point for the 30 year normal annual minimum temperature; calculated for data from 1981 to 2010	LAGOS-US GEO (Smith et al. 2022)

VARIABLE	DESCRIPTION	SOURCE
DEPTH	lake depth (estimated using Hollister method if a measurement was unavailable)	LAGOS-US DEPTH (Stachelek et al. 2021); calculated by GLC for this project
EST_CHLA	Estimated lake chl-a concentration	Brehob et al. 2024
EST_TN	Estimated lake concentration of total nitrogen	Brehob et al. 2024
EST_TP	Estimated lake concentration of total phosphorus	Brehob et al. 2024
FRESHWATER_FISHING_DEMAND	Total number of recreational days per year demanded by people ages 18 and over for freshwater fishing, summarized by HUC-12	EPA EnviroAtlas (USEPA 2016)
GREAT_LAKES_DISTANCE	distance from the Great Lakes (km)	calculated by GLC for this project
HIGHWAY_DISTANCE	straight-line distance to the nearest Class 1 highway (km)	calculated by GLC for this project
INVLAKES_UPSTREAM_MIN_KM	shortest alongstream distance (km) to upstream lake invaded by a spp not yet observed in focal lake	calculated by GLC for this project
INVLAKES_UPSTREAM_N	# of upstream lakes invaded by a spp not yet observed in focal lake	calculated by GLC for this project
LAKE_LAKES1HA_UPSTREAM_HA	total area of lakes greater than or equal to 1 ha upstream of the focal lake, connected via surface streams	LAGOS-US LOCUS (Smith et al. 2021)
LAKE_LAKES1HA_UPSTREAM_N	count of lakes greater than or equal to 1 ha upstream of the focal lake, connected via surface streams	LAGOS-US LOCUS (Smith et al. 2021)
LAKE_NETS_DAMONLAKE_FLAG	A value of 'Y' indicates that there is at least one dam on this lake. This means that the dam point falls onto one of the artificial flowlines that flows through a lake and is therefore associated with the lake and not a stream reach. An "N" indicates no flag.	LAGOS-US NETWORKS, (King et al. 2021); EPA Region V Inland Sensitivity Atlas
LAKE_RSVR_PROBRSVR	probability that a lake is an artificial reservoir	LAGOS-US RESERVOIR
LATITUDE	lake centroid latitude	calculated by GLC for this project
LOG_AREA	base 10 log of area	calculated from AREA
NATURESERVE_LCI	mean NatureServe landscape condition index value of lake watershed	NatureServe Landscape Condition Index (Hak & Comer 2017)
NN_MEAN	mean straight-line distance to nearest obs of each focal AIS in taxon	calculated by GLC for this project
NN_MIN	minimum straight-line distance (km) to a lake with AIS obs	calculated by GLC for this project
NET_LAKES_N	The total number of lakes in the lake network.	LAGOS-US NETWORKS (King et al. 2021)
NLCD_ALLDEV_PCT	percent of zone classified as developed (any intensity) according to NLCD 2016 data	LAGOS-US GEO (Smith et al. 2022)

VARIABLE	DESCRIPTION	SOURCE
NLCD_ALLWET_PCT	percent of zone classified as wetland (any type) according to NLCD 2016 data	LAGOS-US GEO (Smith et al. 2022)
NLCD_CULTCROP82_PCT	percent of zone classified as cultivated crops according to NLCD 2016 data	LAGOS-US GEO (Smith et al. 2022)
NLCD_WETEMERG95_PCT	percent of zone classified as emergent herbaceous wetlands according to NLCD 2016 data	LAGOS-US GEO (Smith et al. 2022)
NLCD_WETWOOD90_PCT	percent of zone classified as woody wetlands according to NLCD 2016 data	LAGOS-US GEO (Smith et al. 2022)
NWS_AREA_HA	area of network watershed	LAGOS-US LOCUS (Smith et al. 2021)
PERCENT_NATURAL	percentage of natural land cover in lake watershed (from NLCD 2021)	NLCD 2021 (Dewitz 2023)
POP_0_4_BUFFER	population within 0.4 km of shoreline	calculated by GLC for this project
POP_5_BUFFER	population within 5 km of shoreline	calculated by GLC for this project
POP_CATCHMENT	population within lake watershed	calculated by GLC for this project
RAMP	presence of at least 1 boat ramp (binary)	EPA Region V Inland Sensitivity Atlas; PA Fish and Boat Commission Access Points (PAFBC 2022); NYSDEC Boat Launches (NYSDEC 2022)
RAMPS_UPSTREAM	# of public boat ramps upstream of focal lake	calculated by GLC for this project
SHORELAND_DIST	Shoreland disturbance - the proportion of the area within 100 m of the lake shoreline with urban or agricultural land cover	calculated by GLC for this project from NLCD 2021 data
SHORELINEDEVFACTOR	Ratio of the lake shoreline to the circumference of a circle of the same area, aka how circular the lake is	LAGOS-US LOCUS (Smith et al. 2021)
STREAMS_ALL_MPERHA	Density of all streams within the lake watershed, calculated as the sum of the stream length divided by the zone area; the artificial lines through non-riverine waterbodies are not counted in the stream length	LAGOS-US GEO (Smith et al. 2022)
WS_AREA_HA	area of lake watershed	LAGOS-US LOCUS (Smith et al. 2021)
WS_NAT_INDEX	Lake watershed naturalness index	calculated by GLC for this project

A number of the included variables were acquired from the data modules of the LAGOS-US Research Platform. Some variables of interest were calculated by the project team because they were not available for the modeled area. Some details of the calculated values are summarized here.

Depth: When a measurement of lake depth was available from the LAGOS-US DEPTH module, that value was used. When no measurement was available, we estimated the lake depth based on the shape of the lake surface and the slope of the surrounding topography, following Hollister et al. (2011). They found an average cross-validated RMSE of 5-6 m depending on geographic region. We were able to process the large number of lakes in the Great Lakes states by doing the calculations in Google Earth Engine using the National Elevation Dataset as the topography data (Figure 4).

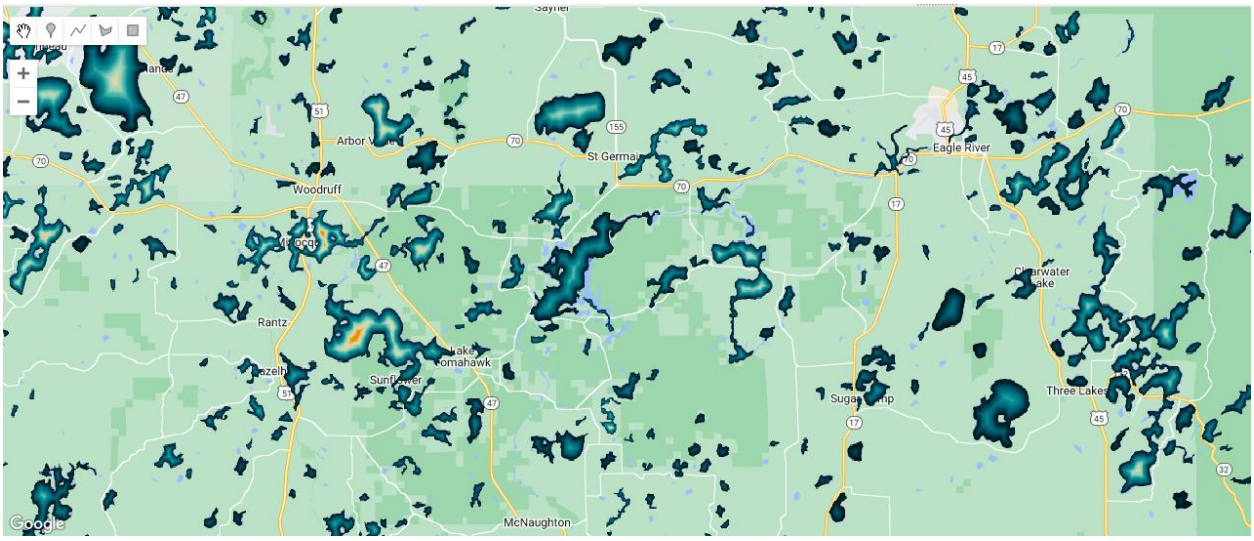

Figure 4: Google Earth Engine screenshot showing lake depth maps calculated from topographic raster data.

Population variables: To better estimate the population living within two buffer distances of each lake (0.4 and 5 km) as well as the population of the lake catchment, we produced a 30 meter dasymetric population raster from 2020 census data using EPA's Intelligent Dasymetric Mapping (IDM) Toolbox. The buffer distances of 0.4 and 5 km represent an "acceptable walking distance" often used in US research studies and the distance within which residents are likely to make regular recreational visits to a lake according to Elliott et al. (2020), respectively.

Ramps: We were able to use an existing inventory of boat access locations developed for EPA Region 5 states. For Pennsylvania and New York, we were able to use the Pennsylvania Fish and Boat Commission's Access Points data layer and the New York State DEC Boat Launches layer, both accessed 6/21/2022. The traffic to/popularity of a boat ramp was estimated using the number of parking spaces at the site as a proxy. Data on the number of parking spaces was available for most ramps, and we filled in missing values by counting parking spaces in aerial imagery.

Lake watershed naturalness index: We quantified the extent of four land cover types (natural, agricultural, low intensity developed, and medium and high intensity developed) in this zone using data from the 2006 NLCD. For each FCN, we calculated the total percentage of floodplain area in each class and then quantified the degree of development using the following weighted index: (1 * % high intensity developed) + (0.75 * % low intensity developed) + (0.25 * % agriculture.) + (0 * % natural). The index ranged from 0 for a floodplain in completely natural cover to 100 for a completely developed floodplain. This index was developed by The Nature Conservancy.

2.2 Data processing

Observations were also spatially clustered near urban areas, especially the metro Detroit and Finger Lakes areas (Figure 3). All of the observation data was used to calculate Because our unit of analysis was individual lakes, we only included the first observation of a species in a given lake in the training data, thinning the presence data and reducing spatial autocorrelation.

The final datasets had a balanced distribution of presences and pseudo-absences, i.e. a sampling prevalence of 50% (Barbet-Massin et al. 2012).

2.3 Validation

To optimize the model's hyperparameters, we applied a time-aware cross-validation that splits the data into temporal folds. This allows us to evaluate our XGBoost model using walk-forward validation, where the model is repeatedly fit on the past data and evaluated. Using temporal folds ensures that the model is evaluated on data that comes chronologically after the training data, mimicking a real-world scenario where future data is not available during training. This helps to assess the model's ability to forecast new AIS introductions in a time series context.

2.4 Algorithm and hyperparameter tuning

Modeling was done in the Python environment using the xgboost package and the scikit-learn open source machine learning library (Pedregosa et al. 2011). Gradient tree boosting, a type of classification and regression tree algorithm, was implemented using XGBoost to identify the most important predictors of AIS observation. Classification and regression trees use dominant patterns in predictor variables to partition the response data into homogenous nodes, or leaves, using a series of decisions, i.e., a decision tree. The resulting decision tree is then used to predict a new series of response data (Breimen 2017).

XGBoost, a machine learning method for regression and classification developed by Chen and Guestrin (2016), was employed because it inherently models variable interactions and seamlessly deals with multicollinearity, both of which are common features of ecological datasets. Unlike traditional statistical methods that start with a data model, boosted regression tree algorithms assume that the process by which the response data were produced is complex and may have resulted from many inherent variable interactions (Elith et al. 2008). Additionally, XGBoost runs more quickly than many other gradient boosting algorithms due to a regularization factor and parallel computing functionality.

A Poisson negative log likelihood loss objective function was used to create a prediction rule for each taxon dataset (fish, plants and invertebrates). Each XGBoost prediction rule was trained with 5-fold temporal cross-validation to identify the number of trees that minimized the objective function. The hyperparameters (the parameters of a machine learning model that aren't optimized in the training process) were optimized using the grid search approach.

2.5 Test data

It is common practice in machine learning applications to test the algorithms on data that were entirely withheld during the processing, tuning, and training steps. The classification performance on withheld test data is used to provide objective measures of model reliability. Given that our objective was the prediction of new AIS introductions, we withheld the observation data for the years 2023 and 2024 and used those records to test the final models' performance. A random subset of the pseudo-absences was assigned to the test data such that a balanced dataset was obtained. This yielded a test dataset with 156, 10, and 407 presence points for plants, fish, and invertebrates, respectively, with matching numbers of pseudo-absence lakes.

2.6 Prediction and mapping

The trained and tuned models were used to generate a continuous score from 0 to 1 for the probability of new AIS detection at every inland lake in the dataset. These scores were exported in tabular and GIS formats for use by end users.

3. Results and Discussion

3.1 "Hotspot" lakes for inland AIS introductions

Early detection hotspots, i.e., areas with a high probability of presence of at least one of the modeled invasive species, tended to be more similar for plants and invertebrates than for fish ().

3.2 Variable contributions

For plants (Figure 5), latitude turned out to be a significant predictor due to the dense spatial clustering of the plant observation data. This was followed by mean temperature, then the minimum and mean distances to the nearest known observation of each invasive species in that model group. Invaded lakes tended to have larger watersheds and more developed shorelines, and recreational boats also played a significant role (presence of a boat ramp, incoming angler connectivity from the FishBrain model, freshwater fishing demand). Overall, the spatial distribution of the available invasive plant observations played the largest role in this model, followed by disturbance and propagule pressure via recreational boating.

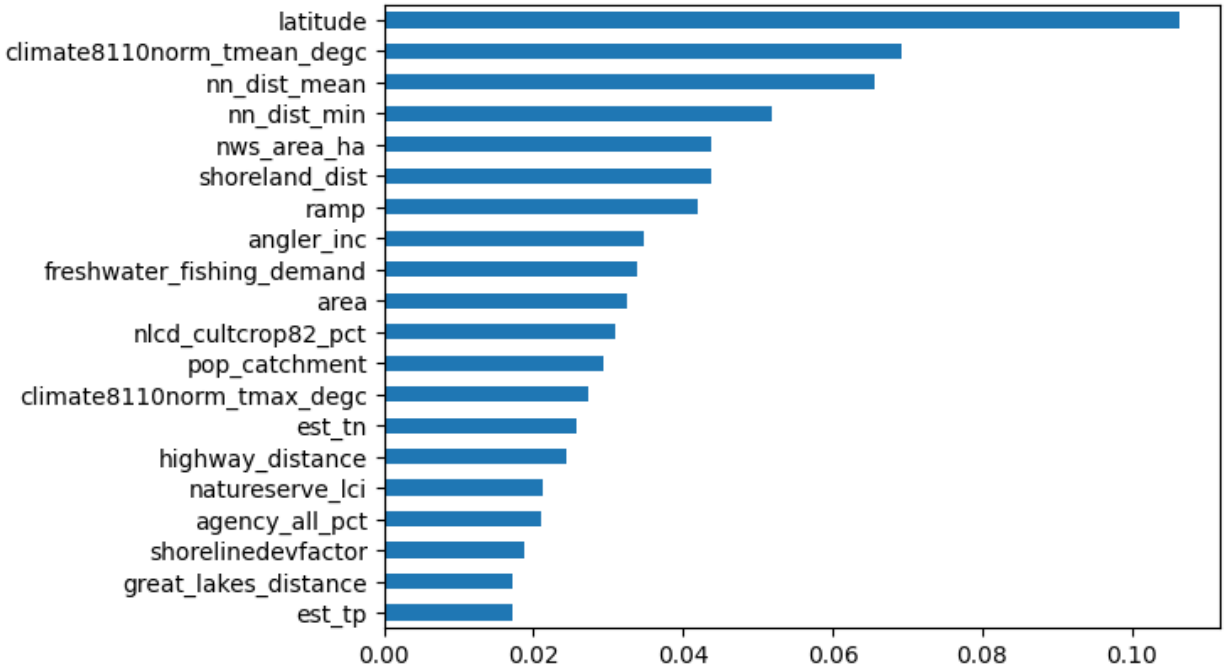

Figure 5: Feature importance values for the 20 most valuable variables in the plant model.

For invertebrates (Figure 6), the presence of a boat ramp at a lake was singularly important, followed by the popularity/traffic at that boat ramp as indicated by the number of parking spaces at the site (boat_parking) and the presence of a ramp upstream from the lake. Other factors such as nearest neighbor distances and lake area played smaller roles.

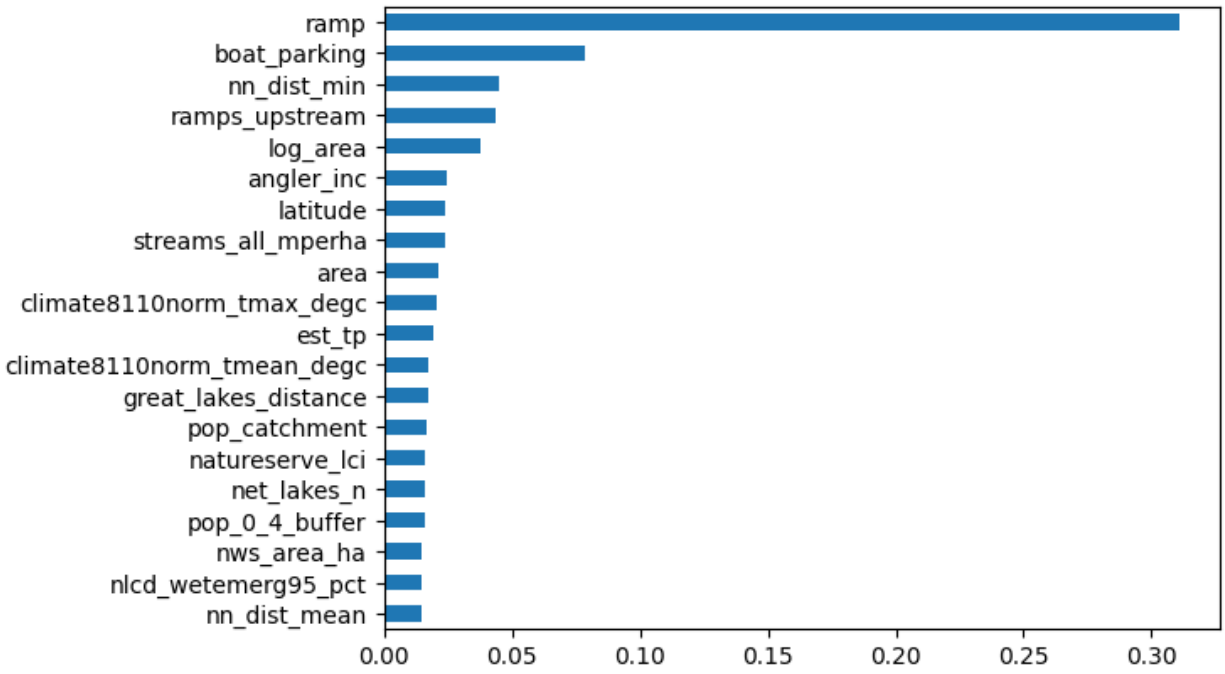

Figure 6: Feature importance values for the 20 most valuable variables in the invertebrate model.

For fish, it was connectivity to other lakes that had the biggest impact on invasion risk, along with minimum nearest neighbor distance again. Lake area played a larger role than for plants and invertebrates. Temperature variables, the probability that a lake was a reservoir, and distance of the lake from the Great Lakes were also meaningful.

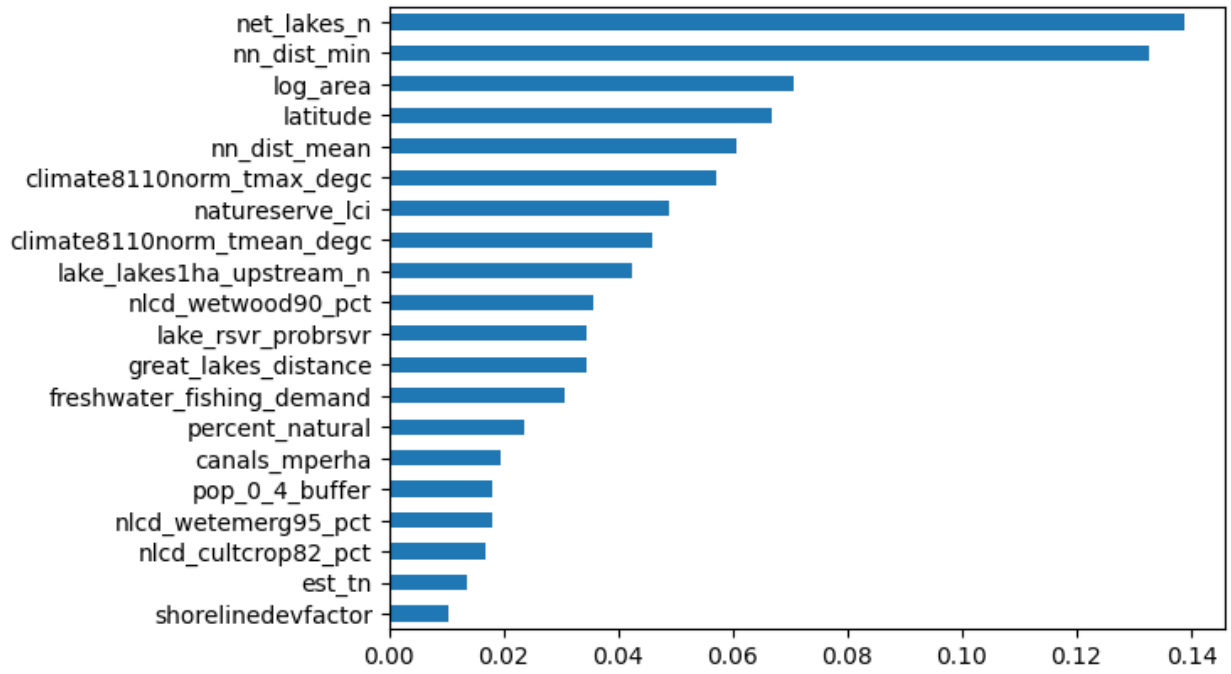

Figure 7: Feature importance values for the 20 most valuable variables in the fish model.

The three models were surprisingly uncorrelated to one another, with pairwise correlation coefficients under 0.2.

3.3 Model performance

All three models performed well, with high ROC AUCs and fairly low cross-validation errors. When used to predict probability of observation at the withheld lakes observed in 2023 and 2024, they continued to perform nearly as well, although we only had 10 observations in the dataset for fish, seriously limiting our ability to evaluate future predictive capacity for that model.

	PLANTS	INVERTEBRATES	FISH
AUC	0.87	0.87	0.87
RMSE	0.164	0.213	0.150
ACCURACY ON TRAINING DATA	0.99	0.99	0.97
ACCURACY ON VALIDATION DATA	0.94	0.94	0.82
2023-24 TEST OBS	156	407	10
ACCURACY ON 2023-24 TEST DATA	0.873	0.923	1.0

3.5 Data visualization

The model results for all three taxa were made available through a map-focused web application developed with ArcGIS. Users can browse results, filter by state, and download the data for their own use as a CSV, shapefile, KML, or geoJSON file (Figure 8). The app can be accessed at https://www.glc.org/work/iedrr/inlandlakesmaps.

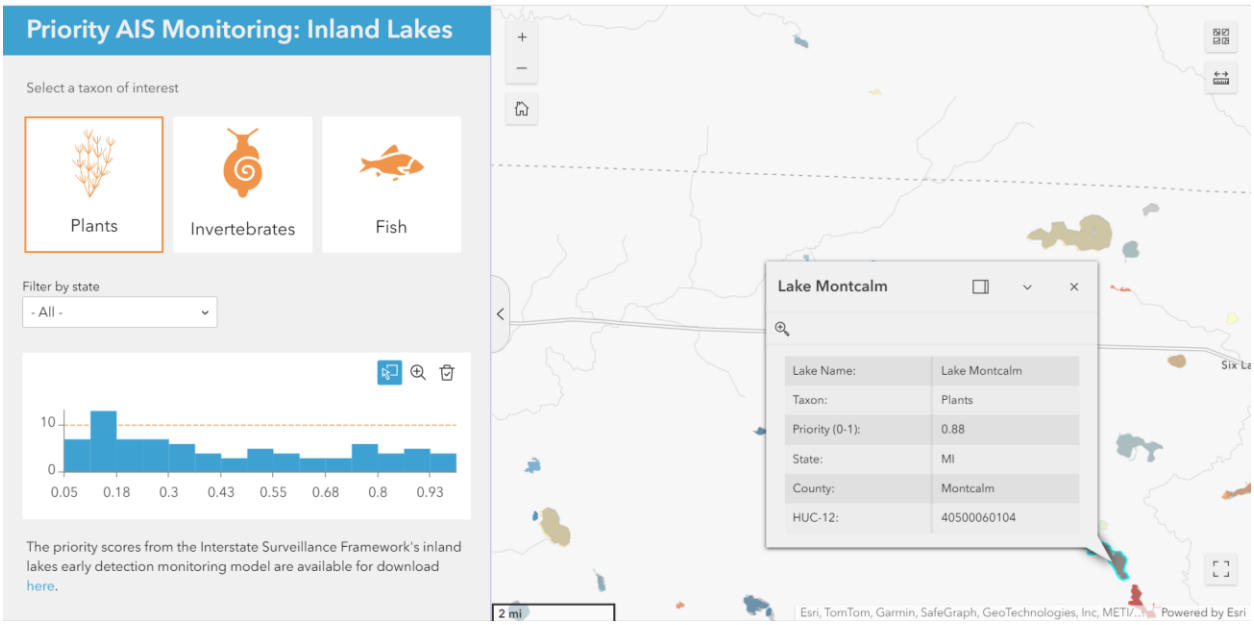

Figure 8: Screenshot of inland lakes model web app.

4. Conclusions and next steps

The strong performance of these models indicates that machine learning can produce hotspot maps for AIS introductions with a high predictive accuracy to help guide prioritization of monitoring efforts. However, the lack of systematically surveyed species presence and true absence data restricted us from disentangling the extent to which predictions were a result of monitoring bias vs. lake characteristics that indeed promote the introduction of new AIS. In Phase 4.5, we will make use of the limited available absence data to further evaluate and improve these maps. We will also continue to investigate the reasons for the lack of correlation across models, and further develop the system to allow users to tailor the results more based on their own program priorities.

5. References

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in ecology and evolution*, *3*(2), 327-338.

Bazzichetto, M., Malavasi, M., Bartak, V., Acosta, A. T. R., Rocchini, D., & Carranza, M. L. (2018). Plant invasion risk: A quest for invasive species distribution modelling in managing protected areas. *Ecological Indicators*, *95*, 311-319.

Brehob, M. M., Pennino, M. J., Handler, A. M., Compton, J. E., Lee, S. S., & Sabo, R. D. (2024). Estimates of lake nitrogen, phosphorus, and chlorophyll-a concentrations to characterize harmful algal bloom risk across the United States. *Earth's Future*, *12*(8), e2024EF004493.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Chamberlain, S. A., & Boettiger, C. (2017). *R Python, and Ruby clients for GBIF species occurrence data* (No. e3304v1). PeerJ Preprints.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Dewitz, J. (2023). National Land Cover Database (NLCD) 2021 Products [Data set]. US Geological Survey. *https://doi. org/10.5066/P9JZ7AO3*.

EDDMapS. (2025). Early Detection & Distribution Mapping System. The University of Georgia - Center for Invasive Species and Ecosystem Health. Available online at http://www.eddmaps.org/; last accessed January 17, 2025.

Edwards, J. L. (2004). Research and societal benefits of the Global Biodiversity Information Facility. *BioScience*, *54*(6), 485-486.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology*, *77*(4), 802-813.

Elliott, L. R., White, M. P., Grellier, J., Garrett, J. K., Cirach, M., Wheeler, B. W., ... & Fleming, L. E. (2020). Research Note: Residential distance and recreational visits to coastal and inland blue spaces in eighteen countries. *Landscape and Urban Planning*, *198*, 103800.

GBIF.org. (2023). GBIF Occurrence Download https://doi.org/10.15468/dl.c75237.

Great Lakes St. Lawrence Governors & Premiers (GSGP). (2019). “Least wanted” AIS list. https://gsgp.org/media/wetfvk5v/least-wanted-ais-brief-5-2019.pdf.

Hak, J.C. and Comer, P.J. (2017), Modeling landscape condition for biodiversity assessment - Application in temperate North America. https://doi.org/10.1016/j.ecolind.2017.06.049.

Higman, P. J., Slaughter, B., Campbell, S., & Schools, E. (2010). Early detection of emerging aquatic and wetland invasive plants in Michigan.

Hollister, J. W., Milstead, W. B., & Urrutia, M. A. (2011). Predicting maximum lake depth from surrounding topography. *PLoS One*, *6*(9), e25764.

Keller, R.P., Frang, K., and Lodge, D.M. (2008) Preventing the spread of invasive species: economic benefits of intervention guided by ecological predictions. *Conserv Biol* 22:80–88.

King, K.B., Q. Wang, L.K. Rodriguez, M. Haite, L. Danila, P. Tan, J. Zhou, & K.S. Cheruvelil. (2021). LAGOS-US NETWORKS v1.0: Data module of surface water networks characterizing connections among lakes, streams, and rivers in the conterminous U.S ver 1. Environmental Data Initiative. https://doi.org/10.6073/pasta/98c9f11df55958065985c3e84a4fe995.

Kinsley, A. C., Haight, R. G., Snellgrove, N., Muellner, P., Muellner, U., Duhr, M., & Phelps, N. B. (2022). AIS explorer: Prioritization for watercraft inspections-A decision-support tool for aquatic invasive species management. *Journal of Environmental Management*, *314*, 115037.

Lozano, V., Di Febbraro, M., Brundu, G., Carranza, M.L., Alessandrini, A., Ardenghi, N.M.G., Barni, E., Bedini, G., Celesti-Grapow, L., Cianfaglione, K. and Cogoni, A. (2023). Plant invasion risk inside and outside protected areas: Propagule pressure, abiotic and biotic factors definitively matter. *Science of the Total Environment*, *877*, 162993.

Milanesi, P., Mori, E., & Menchetti, M. (2020). Observer-oriented approach improves species distribution models from citizen science data. *Ecology and Evolution*, *10*(21), 12104-12114.

MISIN – preferred citation??

NOAA. (2024). GLANSIS Species List Generator. https://www.glerl.noaa.gov/glansis/nisListGen.php.

New York Department of Environmental Conservation (NYDEC). (2022). Boat Launches [Data set]. *https://data.gis.ny.gov/datasets/nysdec::boat-launches/about*. Accessed 6/21/2022.

NYNHP. (2022). Aquatic invasive species pond and lake vulnerability prioritization for New York. https://www.nynhp.org/projects/aquatic-invasive-prioritization/.

Olivero-Sheldon, A., Jospe, A., & Anderson, M.G. (2014). Northeast lake and pond classification system. The Nature Conservancy, Eastern Conservation Science, Boston (MA): Eastern Regional Office.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pennsylvania Fish and Boat Commission (PAFBC). (2022). Access points [Data set]. US Geological Survey. *https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=984*. Accessed 6/21/2022.

Robertson, D. M., & Saad, D. A. (2019). *Spatially referenced models of streamflow and nitrogen, phosphorus, and suspended-sediment loads in streams of the midwestern United States* (No. 2019-5114). US Geological Survey.

Rosaen, A. L., Grover, E. A., & Spencer, C. W. (2016). *The costs of aquatic invasive species to Great Lakes states*. Anderson Economic Group. https://www.andersoneconomicgroup.com/Portals/0/upload/AEG%20-%20AIS%20Impact_%209-20-2016%20Public%20new.pdf.

Smith, N.J., K.E. Webster, L.K. Rodriguez, K.S. Cheruvelil, & P.A. Soranno. (2021). LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. ver 1. Environmental Data Initiative. https://doi.org/10.6073/pasta/e5c2fb8d77467d3f03de4667ac2173ca.

Smith, N.J., K.E. Webster, L.K. Rodriguez, K.S. Cheruvelil, and P.A. Soranno. (2022). LAGOS-US GEO v1.0: Data module of lake geospatial ecological context at multiple spatial and temporal scales in the conterminous U.S. ver 2. Environmental Data Initiative. https://doi.org/10.6073/pasta/53ae2afd051a6a082a2ab129e4281e13

Stachelek, J., L.K. Rodriguez, J. Díaz Vázquez, A. Hawkins, E. Phillips, A. Shoffner, I.M. McCullough, K.B. King, J. Namovich, L.A. Egedy, M. Haite, P.J. Hanly, K.E. Webster, K.S. Cheruvelil, & P.A. Soranno. (2021). LAGOS-US DEPTH v1.0: Data module of observed maximum and mean lake depths for a subset of lakes in the conterminous U.S. ver 1. Environmental Data Initiative. https://doi.org/10.6073/pasta/64ddc4d04661d9aef4bd702dc5d8984f.

Sturtevant, R., Lower, E., Boucher, N., Bartos, A., Mason, D., Rutherford, E., ... & Elgin, A. (2022). A Gap Analysis of the State of Knowledge of Great Lakes Nonindigenous Species. NOAA Technical Memorandum GLERL-176. https://doi.org/10.25923/96kg-kj92.

Valayamkunnath, P., Barlage, M., Chen, F., Gochis, D. J., & Franz, K. J. (2020). Mapping of 30-meter resolution tile-drained croplands using a geospatial modeling approach. *Scientific data*, *7*(1), 257.

Vander Zanden, M.J., Hansen, G.J.A., Higgins, S.N., and Kornis, M.S. (2010). A pound of prevention, plus a pound of cure: early detection and eradication of invasive species in the Laurentian Great Lakes. J Great Lakes Res 36:199–205.

Weir, J. L., Vacura, K., Bagga, J., Berland, A., Hyder, K., Skov, C., ... & Venturelli, P. A. (2022). Big data from a popular app reveals that fishing creates superhighways for aquatic invaders. *PNAS nexus*, *1*(3), pgac075.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ... & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744-751.

United States Environmental Protection Agency (USEPA). (2016). EnviroAtlas. Freshwater Fishing Recreation Demand by 12-Digit HUC in the Conterminous United States. Retrieved: 9/1/2023, from epa.gov/enviroatlas.