https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Mapping Wetland Probability Across Massachusetts with Machine Learning and Multiscale Predictors

Rafter S. Ferguson[1], Sebastian Gutwein[1], Eric Giordano[1], Keith Zaltzberg-Drezdahl[1]

[1] Regenerative Design Group, Greenfield, 01301, United States

Correspondence to: Rafter S. Ferguson (rafterf@regenerativedesigngroup.com)

Abstract. Wetlands perform a vital array of ecosystem functions, but up to 50% of global wetlands have been lost and those that remain are under ongoing threat from development pressures. Accurate and comprehensive maps are critical for the management and protection of wetland resources. Conventional methods for wetland mapping are time consuming and resource intensive, and the common mapping methods that rely on the inspection of aerial imagery often miss forested and other wetland types that do not have a distinctive visual signature, i.e. cryptic wetlands. The use of machine learning and spatial data to map wetlands is a growing field that promises a fast and efficient complement to conventional methods and improved detection of forested and other cryptic wetlands. In this paper we demonstrate the use of a random forest model to generate a large-scale, state-wide map of wetland probabilities in the Commonwealth of Massachusetts, using widely available open source software and publicly accessible data. Through this model we also test the efficacy of multi-scale predictors, including not only terrain derivatives used in previous research but also multi-scale implementations of soil, vegetation, and spectral data. The random forest was trained on the official Massachusetts wetland inventory, and achieved an overall accuracy rate of 92% relative to that dataset. The model showed particular promise in detecting cryptic wetlands by identifying an additional 40% of probable wetland area statewide, and an additional 46% of forested wetland specifically. The use of diverse multi-scale predictors was supported by model performance, variable importance measures, and the feature selection process. This strategy for improving detection of cryptic wetlands and creating better estimates of wetland extent, using non-proprietary software and data, will be a vital adjunct to conventional methods for wetland mapping and monitoring.

1 Introduction

Wetlands perform an array of vital ecosystem functions, supporting biodiversity, regulating and stabilizing water flow, removing pollutants from surface waters, and sequestering carbon (Mitsch and Gosselink, 2015). In many areas, much of historic wetland area has been lost to development, and ongoing development continues to encroach further into wetlands. Global wetland loss since 1900 may be as high as 71% (Davidson, 2014), and the US is estimated to have lost 39% of its historic wetlands (Fluet-Chouinard et al., 2023). The protection of wetlands becomes increasingly important in the face of

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

climate change, as wetlands represent some of the most intense concentrations of sequestered carbon in the biosphere (Nahlik and Fennessy, 2016) as well as providing critical buffers against extreme weather events. Keeping sequestered carbon in wetlands and out of the atmosphere requires vigorous protection from persistent development pressures.

Accurate and comprehensive maps are critical for the management and protection of wetland resources. Development of wetland inventories, however, is time-consuming and expensive, and wetland mapping efforts are constrained by limitations on the capacity of state agencies and non-governmental organizations. Recent advances in remote sensing and machine learning provide techniques and data sets for the rapid and efficient identification of wetlands and can complement and amplify investments in time-intensive field mapping and even conventional desk methods for wetland identification.

The use of machine learning to map wetlands is a dynamic and fast-growing topic in research and planning at the state, national, and international level (Felton et al., 2019; Gale, 2021; Halabisky et al., 2023; Rapinel et al., 2023). Machine learning offers the opportunity to leverage existing datasets to generate or improve assessments of the location and extent of wetlands across large areas, quickly and efficiently. Not intended to replace other methods of wetland identification, machine learning can provide estimates of cryptic wetland area and help identify priority areas for closer inspection, including field inspection and delineation.

In this paper we discuss an implementation of the machine learning algorithm random forest to produce a wetland probability map of the Commonwealth of Massachusetts. In order to estimate the extent of cryptic wetlands that do not appear on existing inventories, we developed a machine learning model making use of statewide spatial data and widely available open-source software. The wetland probability map can be combined with existing inventories to create a more accurate and inclusive estimate of total wetland area.

1.1 Digital wetland mapping with random forest

The growing field of digital wetland mapping has drawn on an array of machine learning algorithms. Any suitable algorithm must be able to handle large and complex datasets while making few assumptions about the distribution of variables. Random forest is an ensemble learning algorithm that meets these criteria, and is increasingly recognized as one of the most powerful and effective machine learning strategies for wetland detection (Jafarzadeh et al., 2022). Random forest creates an ensemble of decision trees, each trained on a random bootstrap sample of the data and a random subset of predictors (Breiman, 2001). The model prediction is the averaged prediction of all individual trees, with the prediction appearing as a probability value between 0 and 1. In addition to meeting the criteria outlined above, the bootstrapping and aggregating strategy reduces bias, avoids overfitting, accommodates correlated predictors, has been shown to produce highly accurate predictions without extensive tuning of hyperparameters, i.e. “out of the box,” with default settings, accounts for complex

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

multi-way interactions, and has a built-in variable importance mechanism that is useful for feature selection–all of which have led to extensive use of random forest in remote sensing (Belgiu and Drăguţ, 2016). Recent years have seen growing use of the random forest algorithm in state and regional wetland detection projects in the US (Maxwell et al., 2016; O'Neil et al., 2018) and incorporation into a toolbox in popular proprietary GIS software (The Wetland Identification Model (WIM) – A New Arc Hydro Functionality for Predicting Wetland Locations Using LiDAR Elevation Data and Machine Learning, 2024).

Digital mapping of wetlands can draw on a wide range of potential predictors. The presence of wetlands in the landscape is determined by the interaction of terrain and hydrology, and is in turn reflected in a range of soils, vegetation, hydrological characteristics, and spectral signatures, all of which have been used as inputs in various wetland mapping projects (Jafarzadeh et al., 2022). For all predictors, spatial scale is a key question. Wetlands are produced by the interaction of forces operating and combining across multiple spatial scales, and there is no clear cut method for determining which scale is most relevant across contexts. Choice of spatial scale for predictors is especially pertinent for approaches like random forest, which uses point-based intersections of predictors and does not natively account for contextual information. Using multi-scale predictors, i.e. calculating, resampling, or smoothing spatial predictors at multiple scales, can therefore help address several challenges at once. In the case of terrain derivatives, it is clear that topography can control the accumulation and retention of water at the scale of local undulations in grade (e.g. 1-10 meter) or broader topographic depressions (e.g. 10-100 meter, or larger). Other classes of predictors, such as vegetation height, soil type, or spectral data, can show similar multi-scale neighborhood effects, where what matters at a given sample point is not the value of the predictor at that point, but the average value in the area that forms its context. In the case of soil-based predictors, multi-scale smoothing can also help by creating fuzzy boundaries between soil units, which has the potential to ameliorate some of the known challenges with the crisp and sometimes arbitrary delineations that soil maps impose on underlying heterogeneity and varying transition zones (Hunter et al., 2009; Nikiforova et al., 2020). As there is no way to know a priori which scales will be relevant for which predictors, the variable importance scores generated by the random forest algorithm can be used to select the most relevant predictors and scales.

In the related domain of digital soil mapping, the use of such multi-scale terrain derivatives is becoming increasingly well established (Behrens et al., 2019), and has been shown to increase prediction accuracy dramatically (Miller et al., 2015). Application to other classes of predictors aside from terrain, however, remains somewhat neglected. Recent work has introduced the use of multi-scale terrain features specifically for digital wetland mapping (Halabisky et al., 2022).

1.3 Study Area

Massachusetts is a coastal state in the northeastern United States with a diverse range of soils, topography, and vegetation. While Massachusetts was among the first states to implement wetland protections in the mid-20th century, it's estimated

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Massachusetts has lost a third of its wetlands since colonial times (Protecting Wetlands in Massachusetts | Mass.gov, 2024). Roughly 30% of the state by area, in the western and north-central zones, is in the EPA Level III Ecoregion the Northeastern Highlands, characterized by high elevations, rugged topography, mixed deciduous and coniferous forests, and soils built on glacial till (Griffith et al., 2009). From the central through the eastern coastal zone, the Northeastern Coastal Zone makes up roughly 60% of the commonwealth. The Northeastern Coastal Zone is marked by lower elevations, less topographical variation, and a prevalence of sandy loams inland and sandy soils with low organic content in coastal areas. The remaining 9% of the state is in the Atlantic Coastal Pine Barrens, a low-lying area with diverse coastal landforms including dunes, marshes, and bays, with vegetation dominated by scrubby oak-pine forests.

1.4 Objectives

Our principal objective was to demonstrate a method for generating a large-scale, state-wide map of wetland probabilities, using widely available open source software and publicly available data. Such a method could serve as a fast and efficient adjunct to conventional field and desk methods, and in turn support many goals for wetland monitoring, management, and conservation. A secondary objective was to use machine learning to leverage the utility of an existing wetland inventory by using it as training data that (together with the array of predictors) could identify additional probable wetland area missing from the same inventory. Our tertiary objective was to assess the multi-scale predictor approach using a broader swath of environmental variables than has previously been demonstrated in wetland detection efforts.

2 Methods

The data and modelling processes are represented in Fig. 1. All operations in this project were performed with free open source software. Acquisition and preprocessing of geospatial data were performed using QGIS (QGIS Development Team, 2023), and data wrangling and modelling were conducted in the R statistical programming language (R Core Team, 2023).

2.2 Data

The Massachusetts Department of Environmental Protection (MassDEP) has produced an official state wetland inventory that represents a distinctively high-quality resource of state-level wetland information for planning and research in Massachusetts (Baker et al., 2019). The first version of the MassDEP Wetlands map was completed in 2006, based on visual assessment of 1:12,000 stereo imagery taken between 1990 and 2000. This original map featured field validation for up to 10% of identified wetlands. Starting in 2007, MassDEP created an updated layer, using 2005 imagery at 1:5,000 scale. Production of the 2005 MassDEP Wetlands layers (MDW) did not feature any field validation. The layer compares favorably with the National Wetland Inventory (NWI), which is produced with similar methods but using older and larger-scale imagery. The present utility of the MDW is constrained by both the intrinsic limitations of the method that it shares with the NWI, and with the age of the imagery used to produce it. Numerous studies have confirmed that the NWI misses many

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

wetlands, particularly but not only forested and smaller wetlands (each of which may in aggregate make up a significant portion of total wetlands) (Kudray and Gale, 2000; Martin et al., 2012; Matthews et al., 2016; Stolt and Baker, 1995; Tiner, 1990). While no comparison study has been conducted in Massachusetts, reports from field delineators, together with comparison with acquired field delineations, confirm that similar limitations apply to the MDW. Additionally, trends in wetland loss and gain are ongoing and dynamic, and the imagery on which the MDW is based is now nearly two decades old (Baker et al., 2019).

Training and testing data were developed from the MDW. The layer was edited to remove open water according to the DEP Wetland designation. Given the scale of the study area, researchers opted for a moderately dense sampling approach, creating a training dataset of 30,000 points. Reference wetland data were assembled from 10,000 random points drawn from within wetland polygons, and background data were assembled from 20,000 random points not within wetland polygons. Non-wetland points are treated as 'background' or 'pseudo-absence' rather than true absence points, as it is expected that some of the 20,000 background points were placed within cryptic wetlands that do not appear in the MDW. This distinction was accommodated in our downsampling approach described in section **2.3 Model Fitting** below.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

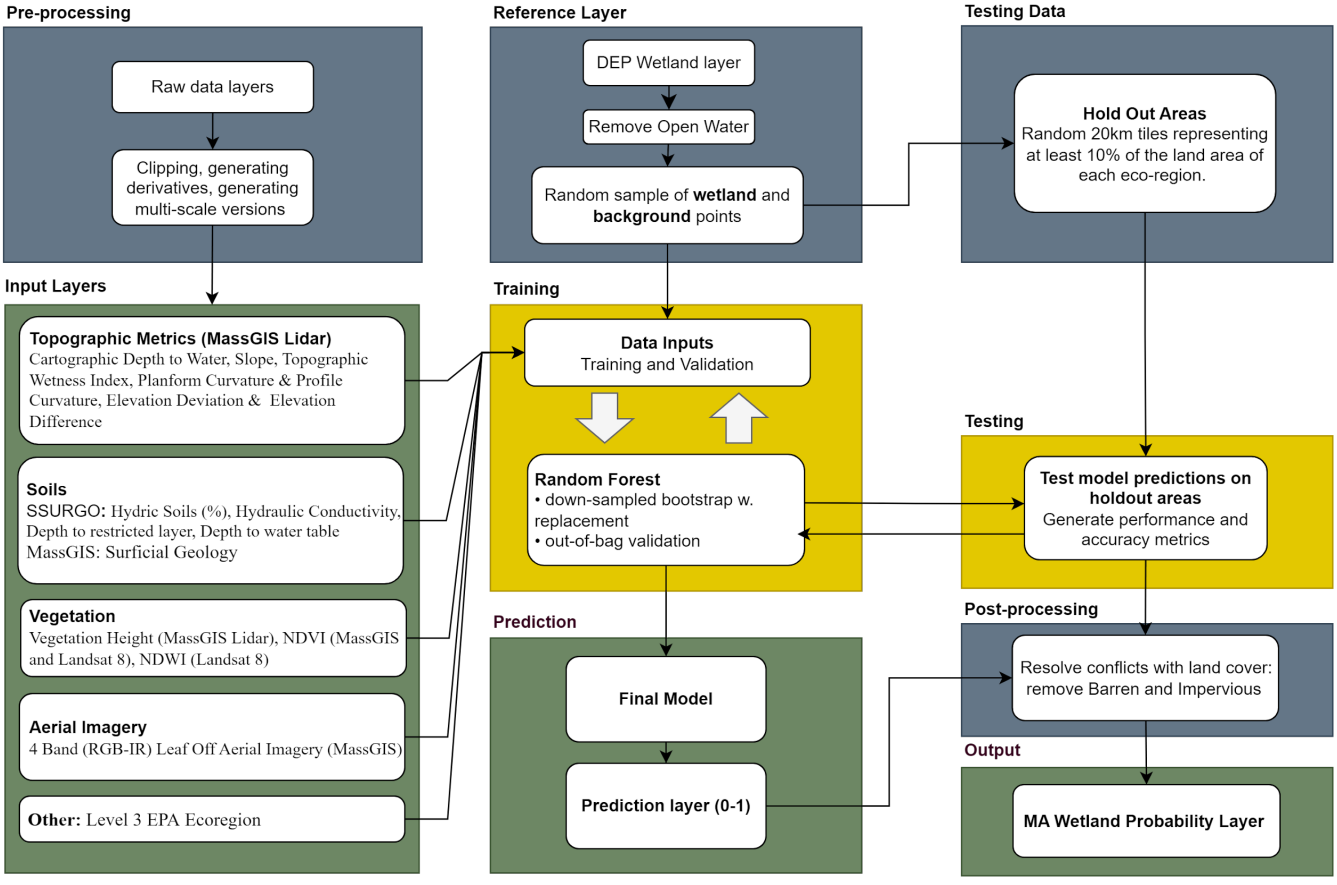

Figure 1. Predictors and process flow for generating a statewide wetland probability layer. Using machine learning to generate a Wetland Probability Score at 4m resolution will support a more inclusive estimation of wetland extent across the state.

We extend the logic of the multi-scale approach in digital soil mapping, and the application of multi-scale terrain derivatives to wetland mapping, by aggregating all predictors to multiple larger resolutions (whenever feasible based on the starting resolution of available layers), and including all resolutions in our initial model. Figure 2 illustrates three of the most important predictors rendered at multiple scales for the same area**.** Approaches to resampling included specification of neighborhood size for terrain derivatives and mean smoothing or filtering for other variables.

Terrain variables were developed from the MassGIS LiDAR point cloud data. The point cloud was used to generate a digital elevation model (DEM) for the entire state, resampled to 4m resolution. The DEM was in turn used to generate derivative

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

measures at multiple scales, including Slope, Cartographic Depth to Water (DTW), Planoform Curvature, Profile Curvature, and Topographic Wetness Index. Cartographic Depth to Water is a metric that estimates the difference in elevation between the soil surface and the closest open water (Murphy et al., 2011). Topographic Wetness Index (TWI) is a predecessor of DTW that is used, like DTW, to model soil moisture conditions (Murphy et al., 2009). Profile and planform curvature quantify the curvature of the earth's surface in the direction of slope, and perpendicular to slope, respectively (Maxwell et al., 2016).

With one exception, soil variables were drawn from the Soil Survey Geographic Database (SSURGO) maintained by the US Natural Resources Conservation Service (Soil Survey Staff, n.d.). These variables included Depth to a Restrictive Soil Layer, Depth to Water Table, Percent Hydric Soils, and Hydraulic Conductivity (Soil Survey Staff, 2015). Depth to Restrictive Soil Layer is the estimated distance from the soil surface to a layer that impedes the movement of water (often bedrock). Percent Hydric Soils refers to the percent of the SSURGO map unit that is composed of hydric components. Hydraulic Conductivity is the volume of water that would move through the soil under saturation conditions in a standardized area and unit of time. The vector-based variables from SSURGO were modified to multiple scales by generating a raster layer for each of the relevant fields of the SSURGO vector layer, at 4m resolution, and then applying a mean filter at 12m and 60m. An additional variable, Surficial Geology, was acquired from the MassGIS Surficial Geology layer, which was in turn acquired from the US Geological Survey (Stone et al., 2018).

Vegetation variables were acquired from multiple sources. The mean and median of NDVI Composite and Normalized Difference Water Index (NDWI) Composite were each acquired, based on 30m resolution LANDSAT 8 imagery, directly via Google Earth Engine. Normalized Difference Vegetation Index (NDVI) was calculated from 2021 MassGIS Leaf-Off Aerial Imagery, and aggregated to multiple scales. NDVI is a widely used metric of vegetation health that uses the red and near-infrared bands to monitor the level of photosynthetic activity. NDWI similarly uses the green and near-infrared bands to assess moisture levels. Vegetation height was calculated based on first returns from the MassGIS LiDAR point cloud data sets, and aggregated to multiple scales. The EPA US Level 3 Ecoregions for Massachusettswere also included as a three-value categorical predictor: Atlantic Coastal Pine Barrens, Northeastern Coastal Zone, and Northeastern Highlands. In addition, bands one through four (red, green, blue, and near infrared, respectively) from 2021 MassGIS Leaf-Off Aerial Imagery were acquired and smoothed to multiple scales.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

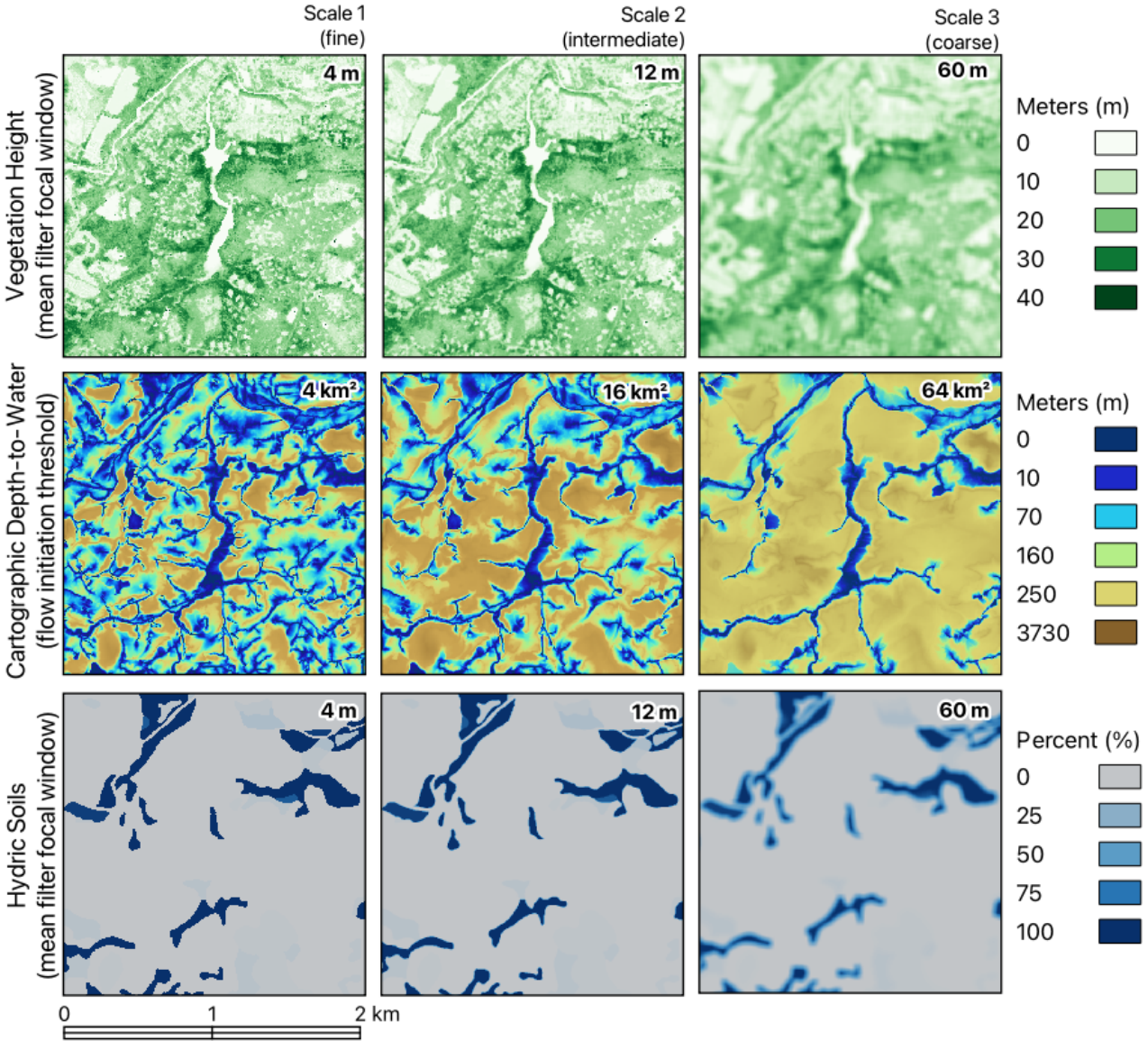

Figure 2. Multi-scale predictors. Three of the predictors included in the final model are shown for the same area, an approximately 2 km square centrally located in the Northeastern Coastal Zone. Predictors are shown across the range of scales that were each included in the model. The method of multi-scaling for each predictor is indicated in parentheses below the predictor name on the left side of the figure. Model predictions for the area shown here can be seen in callout (C) in Fig. 4.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

2.3 Training and testing

Model testing was conducted with two different spatial configurations of holdout data. In the interest of a more rigorous assessment of model performance and generalizability, holdout data for testing the final model was based on the random selection of 20 km by 20 km tiles within each EPA Level 3 ecoregion (Fig. 3). Within each of the three ecoregions in Massachusetts, tiles were randomly selected for holdout until at least 10% of land area within that ecoregion was designated as holdout (Table 1). Points in these tiles were held back from the training data, comprising 23% of the 30,000 points. By using tile-based holdout, predictive performance could be assessed based on nearby but unsampled landscapes, rather than on holdout points interspersed within the same landscapes on which the model was trained. In order to assess the impact of the tile-based holdout on model performance, we fit a model with identical parameters but with testing data withheld fully randomly, ignoring tile structure, comprising 25% of sample points across the state.

Table 1. Holdout areas for model testing

		Holdout area			Holdout points	
Ecoregion	**acres**	**%**	**acres**	**tiles**	**wetland**	**background**
Atlantic Coastal Pine Barrens	488,314 ac	17%	80,647	4	181	293
Northeastern Coastal Zone	3,134,117 ac	26%	813,232	9	1637	2959
Northeastern Highlands	1,567,251 ac	21%	332,372	5	459	1484
Total	5,189,682 ac	21%	1,226,251	18	2277	4736

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

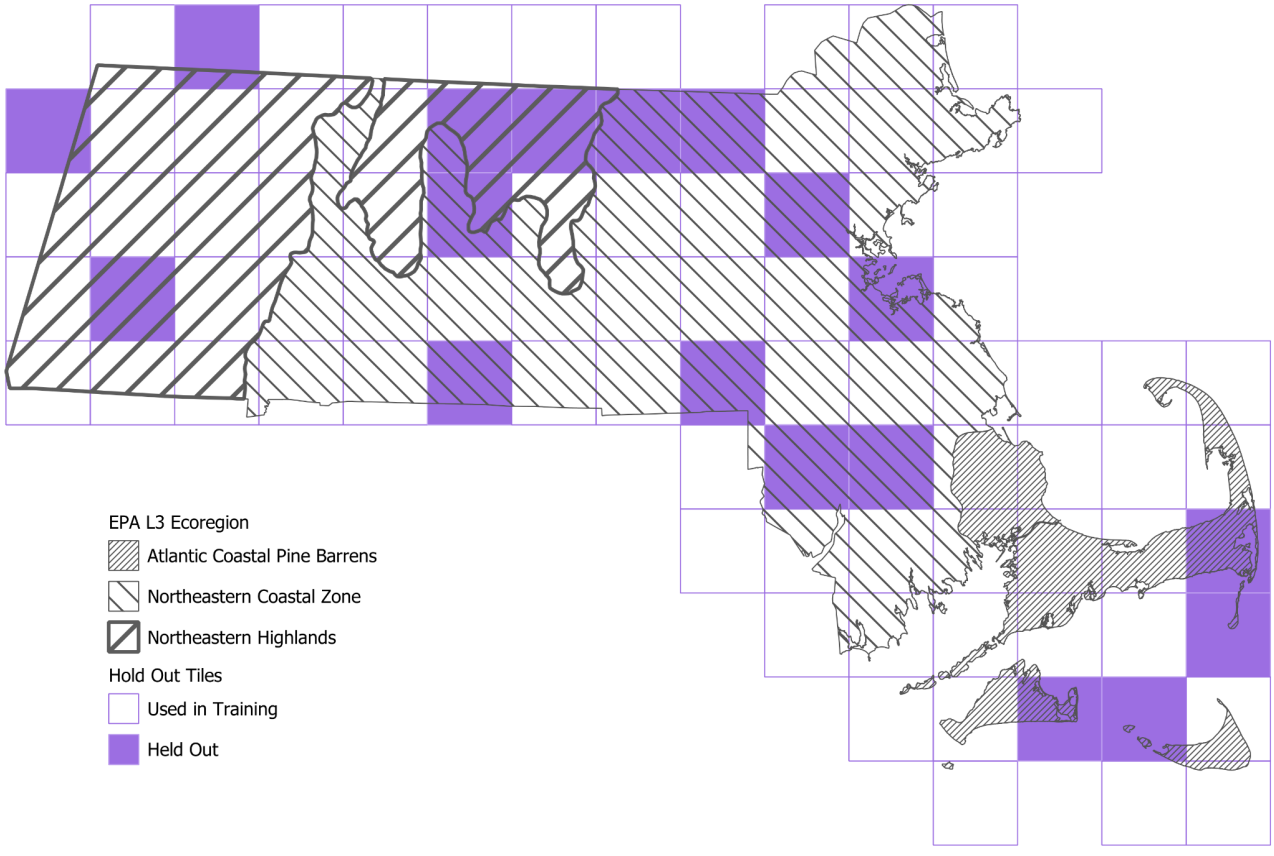

Figure 3. EPA Level 3 Ecoregions for Massachusetts with Hold Out Areas

2.3 Model fitting

The wetland probability score was generated using a random forest (RF) model, using the canonical randomForest package in R (Liaw and Wiener, 2002). Before fitting the statewide model, researchers generated a succession of preliminary models covering progressively larger areas. Preliminary models were used as sanity checks, by examining out-of-bag (OOB, see section **2.4 Model validation and testing** below) metrics, visual inspection of overlays with the DEP wetland map, aerial imagery, and field wetland delineations, and comparison with landscapes known to the researchers.

The first statewide model was fit using all 63 predictors (23 variables, each at 1-4 scales). In order to reduce model complexity and processing time, feature selection was carried out using a holdout variable importance procedure that compares full ensembles fit with and without the predictor, using the randomForestSRC package (Ishwaran and Kogalur, 2007). Through this process 34 predictors were found to be making no contribution to model accuracy. The reduced model

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

was fit with the remaining 29 predictors (Table 2), and compared to the full model to assess any impact on out-of-bag accuracy.

While the more conservative and time-intensive holdout variable importance procedure was used for variable selection, the more widely used permutational variable importance measure was used to understand the relative influence of the 29 variables included in the final model (Belgiu and Drăguţ, 2016). Permutational variable importance compares the OOB prediction accuracy using the original variable against the accuracy with the variable randomly permuted. The variable importance is calculated as the mean decrease in accuracy under permutation. As the stochastic nature of the random forest algorithm makes variable importance rankings shift across multiple runs of the model, nine additional models were fit with identical parameters, and variable importance measures extracted from each, in order to provide a sense of range and stability of the importance measure.

The use of the DEP wetland map as training data meant that the training data lacked true absence points, necessitating the assumption that some of the ostensibly non-wetland points would fall in unmapped, cryptic wetlands. The use of background or pseudo-absence points is a common situation in the domain of species distribution modelling (SDM), which generally assumes that background points do not reliably represent true absence. Random forest is also used extensively in SDM, and researchers in that field have evaluated multiple approaches for dealing with the lack of true absence points. Among the best performing and simplest of these approaches is down sampling, or the downward adjustment of the bootstrap sample to match the number of presence points (Valavi et al., 2021, 2022). The default bootstrap size for the randomForest function used in this study is 0.632 of the number of observations in the training data. Use of the default setting would have produced a bootstrap sample size of 14,438. To implement down sampling, the bootstrap sample of the random forest model was set to 7586, the number of wetland points in the training data. In order to assess any effect of down sampling on model performance, the full model was also fit without down sampling, i.e. with default bootstrap settings.

2.4 Model validation and testing

Model performance was assessed in several ways, using both out-of-bag prediction and test data. Out-of-bag prediction is a strength of the random forest approach, made possible by the bootstrap sampling process. At each iteration, a decision tree is grown using a subsample of the training data. Out-of-bag refers to the data which was not part of a given bootstrap sample, and therefore not used to grow the associated decision tree. Each decision tree can therefore be validated by using it to predict the OOB data, as they have not been used in its training. Averaging the error rate over all trees in the forest gives the OOB error rate for the model. While OOB metrics are sometimes used in place of external holdout testing data, for this project testing with external data was performed as well.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Table 2. Predictors used in the final model

Name	Description	Type	Source Data	Scales	In Final Model	Scale Unit
Depth_ Restrictive	Estimated Depth to a Restrictive Soil Layer	Soil	NRCS SSURGO	4, 12, 60	12	meters
Depth_ Water	Estimated Depth to Mean Water Table	Soil	NRCS SSURGO	4, 12, 60	4, 12	meters
Hydric_ Soils	Percent Hydric Soils in Map Unit	Soil	NRCS SSURGO	4, 12, 60	4, 12, 60	meters
KSat	Estimated Hydraulic Conductivity	Soil	NRCS SSURGO	4, 12, 60	60	meters
Surf_Geo	Surficial Geology	Soil	MassGIS Surficial Geology	4	4	meters
B2	Aerial Imagery: Green Band	Spectral	MA 2021 Leaf Off Aerial Imagery	4, 12, 60	12	meters
B4	Aerial Imagery: Near Infrared Band	Spectral	MA 2021 Leaf Off Aerial Imagery	4, 12, 60	60	meters
DEV	Deviation in Elevation Relative to the Neighborhood Pixels	Terrain	LiDAR Derived Elevation	4, 12, 60, 300, 1000	12, 60	meters
DIF	Difference in Elevation Relative to the Neighborhood Pixels	Terrain	LiDAR Derived Elevation	4, 12, 60, 300, 1000	12, 60	meters
DTW	Cartographic Depth to Water	Terrain	LiDAR Derived Elevation	250, 1000, 4000, 16000	250, 1000, 4000	catchment area (pixels)
Slope	Slope of the Topography in a Neighborhood	Terrain	LiDAR Derived Elevation	4, 12, 60, 300, 1000	12, 60	meters

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

NDVI_ Mean	Mean of the Landsat 8 Normalized Difference Vegetation Index Composite	Vegetation	LANDSAT 8 from Google Earth Engine	30	30	meters
NDVI_ Median	Median of the Landsat 8 Normalized Difference Vegetation Index Composite	Vegetation	LANDSAT 8 from Google Earth Engine	30	30	meters
NDWI_ Mean	Mean of the Landsat 8 Normalized Difference Water Index Composite	Vegetation	LANDSAT 8 from Google Earth Engine	30	30m	meters
Veg_ Height	Vegetation Height	Vegetation	LiDAR Derived Vegetation Height	4, 12, 60	4m, 12m, 60m	meters

As the model produces a 0-1 probability score and not a classification, in order to calculate accuracy metrics it is necessary to collapse the probability score into a dichotomous outcome (i.e. wetland/upland). The standard probability threshold of .5 was used to bin pixels into wetland and non-wetland categories. Predictions can then be compared to the MDW. Model accuracy was explored through several metrics: overall accuracy, and true/false positive and negative rates. It is important to remember that true and false are here relative to the reference layer, and that some amount of over-prediction relative to that layer (i.e. identification of additional wetlands) is a fundamental goal of the project. The confusion matrix captures values for true and false negatives and positives in a table that crosses the values of the reference data with the model predictions (Ting, 2010). Accuracy and confusion matrices were also calculated for alternate parameterizations and sampling strategies in order to understand any impacts of different strategies on model performance.

Additionally, the model was evaluated with a metric designed specifically for probabilistic binary classifiers, which does not require collapse of continuous probability scores into dichotomous outcomes. The measure known as *area under the receiver operating characteristics curve* (AUC) provides a metric that has been found useful for binary classifiers, including in digital wetland mapping (Maxwell et al., 2016). The curve from which the AUC is derived plots the true positive rate against the false positive rate across a spectrum of probability thresholds. For the purposes of interpretation, the AUC is equivalent to the probability that the model will assign a higher probability score to a randomly chosen wetland point than to a randomly chosen background point (Fawcett, 2006). An AUC of 1 indicates perfect prediction, while an AUC of .5 would mean model predictions are totally random. For our purposes, its utility is in generating a single measure of probabilistic model

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

performance. While there are no hard and fast thresholds for AUC scores in remote sensing, scores above 0.70 are often judged acceptable, above .80 good, and above .90 excellent (Hosmer et al., 2013).

In the interest of understanding how the model performed on different wetland types and land covers, DEP wetland types were crosswalked with categories of land cover in the MassGIS 2016 Land Cover/Land Use dataset. The crosswalk was supplemented with coastal wetlands data accessed through the Northeast Oceans Data Portal, which were used to assign predicted additional wetlands to the category of salt marsh (Northeast Ocean Data Portal, 2024). In this way, model predictions of previously mapped DEP wetland classes could be compared with model predictions of additional wetlands that, as they do not appear in the DEP wetland inventory, have not been assigned a wetland type.

In order to assess impact of model predictions on state-level wetland extents, the wetland probability map was post-processed using the MA 2016 High Resolution Land Cover dataset to remove the land cover classes of impervious and barren, which were identified as areas where the prediction was likely misidentifying wetlands. Additionally, the areas identified as Open Water in the MDW were removed.

3 Results

Initial models were fit with 500 trees. After plots of error rate against number of trees showed that the model achieved stability after approximately 100 trees, subsequent models were fit with 400 trees to reduce processing time. Initial models were fit with 63 predictors, and after feature selection using holdout variable importance, a reduced model was fit with 29 predictors. After confirming equivalent performance with the reduced model, the latter was selected as the final model. Model output of the Wetland Probability Score for the entire state is shown in Fig. 4.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

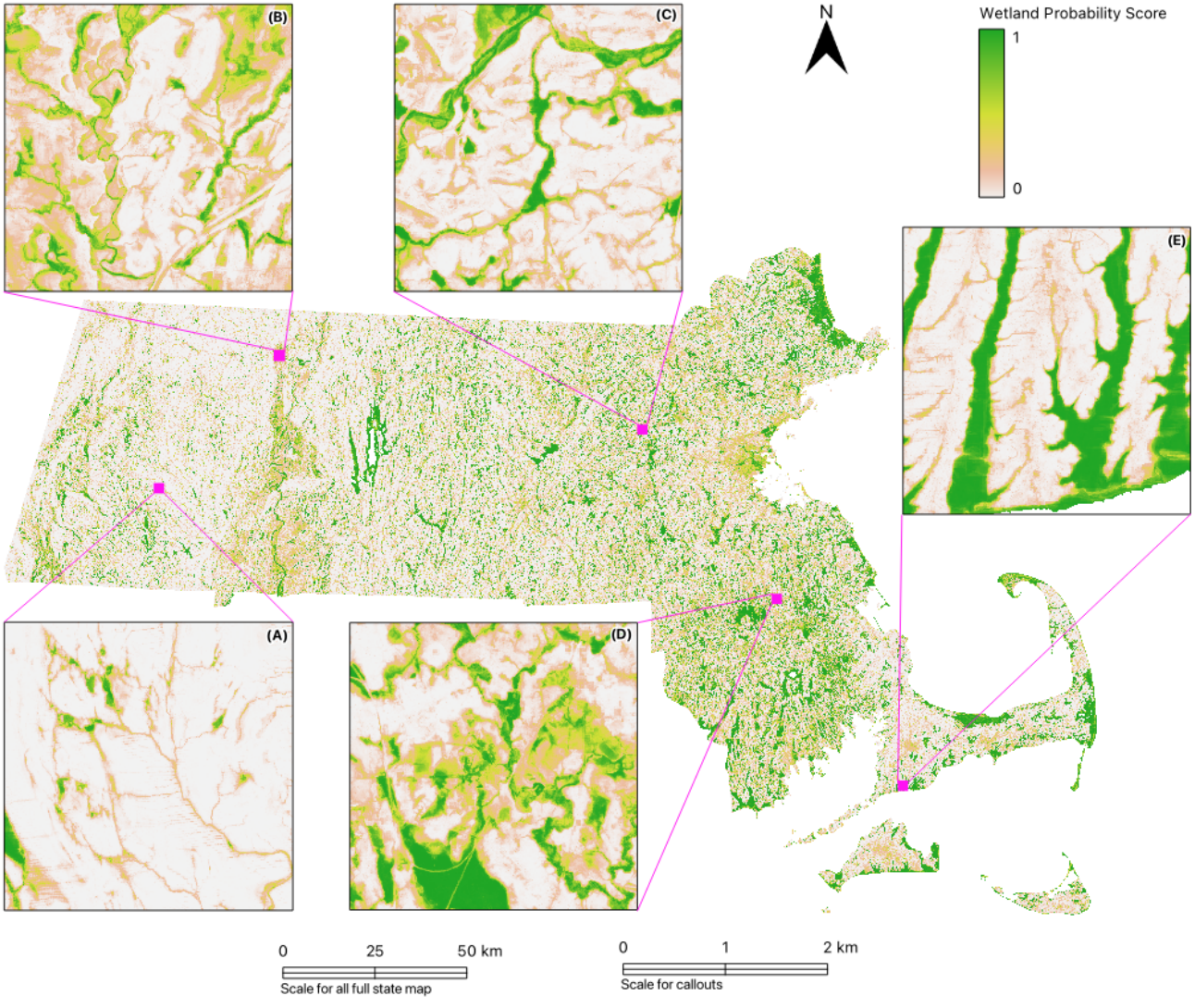

Figure 4. Wetland Probability Score for Massachusetts. This map reflects the raw WPS without post-processing to remove conflicting land use designations, e.g. open water and impermeable surfaces. From west to east, callout (A) is in the Berkshires Hills of the Northeastern Highlands ecoregion; callouts (B-C) are in the Northeastern Coastal Zone, with (B) in the Connecticut River Valley, (C) in the Gulf of Maine Coastal Plain, and (D) in the Narragansett/Bristol Lowland; and callout (E) is in Atlantic Coastal Pine Barrens. Callout (C) shows the same area represented in Fig. 2. Multi-scale predictors.

3.1 Accuracy and Prediction Outcomes

Model performance met or exceeded all expectations for prediction. Generating predictions on the held-out testing data, and a dichotomized prediction (wetland/upland at a probability threshold of .5), the model achieved an overall accuracy rate of

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

92%, with a true positive rate of 87% and a true negative rate of 94%. Prediction accuracy based on the testing data was nearly identical to the out-of-bag accuracy score generated through the model fitting process. Performance was similar among all comparison models (Table 3). The probabilistic metric AUC was 0.971, i.e. there is a 97% chance that the model would assign a higher probability score to a randomly selected wetland point than to a randomly selected non-wetland point. While there are no hard and fast thresholds for acceptability and excellence for AUC scores, by any standard 0.971 indicates outstanding performance (Hosmer et al., 2013).

Table 3. Confusion matrices and performance metrics for final model and comparison models

					Reference data (DEP Wetlands)				**Metrics**	
	Train/test	**Bootstrap**	**Type of test**	**Model prediction**	Counts		Rates		Accuracy	AUC
					0	1	0	1		
Final model 29 predictors	Tile-based holdout	Down sampled	Holdout	0	4476	289	95%	13%	0.922	0.971
				1	259	1977	5%	87%		
			Out-of-bag	0	14459	775	93%	11%	0.918	
				1	1068	6155	7%	89%		
Comparison models 63 predictors	Random holdout	Down sampled	Holdout	0	4770	351	94%	13%	0.919	0.973
				1	244	2010	6%	87%		
			Out-of-bag	0	14210	745	93%	11%	0.918	
				1	1061	6067	7%	89%		
	Tile-based holdout	Default sampling	Out-of-bag	0	14457	790	93%	11%	0.918	
				1	1077	6339	7%	89%		

True Neg	False Neg
False Pos	True Pos

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

3.2 Variable Importance

Terrain and soil variables were the most decisive predictors of wetland presence (Fig. 5). The LIDAR-derived cartographic Depth to Water (DTW), and the estimated percent of hydric soils from SSURGO, at various scales comprised the top six predictors and seven out of the top eight. Examination of the range of importance scores (mean decrease in accuracy) for each predictor, across 10 model fits, shows clusters of stability within rankings that shifted from iteration to iteration. DTW_1k held the top rank in every iteration, while Hydric_Soils_60m and DTW_4k each consistently held either the 2nd or 3rd positions. Future investigations may focus primarily on the highest-performing variables from this predictor set.

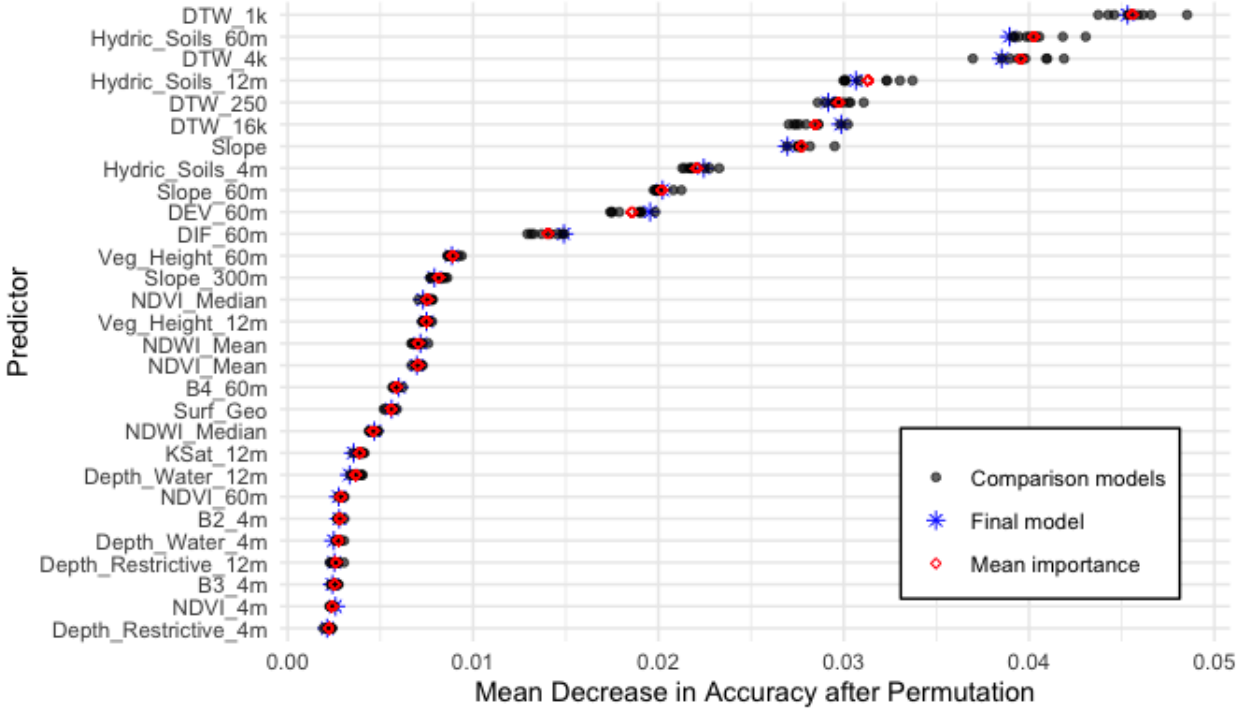

Figure 5. Variable importance in the final model. Permutational variable importance of the 29 variables used in the final model, across 10 models fit with identical parameterization. Plotting importance across multiple models allows for the assessment of relative importance while accounting for stochastic variability across model fittings. The red diamond indicates the mean importance for that variable across all 10 models, and the blue asterisk shows the importance within the final model itself.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Examining the relationship between important variables and prediction outcomes yields some understanding of patterns in over- and under-prediction. Figure 6 shows the distribution of predictor values across prediction outcomes, i.e. true and false positive, and true and false negative, for the three most important multi-scale predictors: cartographic depth to water, percent hydric soils, and slope. DEP wetland and background points show consistent separation across important predictors, with little or no overlap of interquartile ranges between true positives and true negatives. Unsurprisingly, atypical values for these predictors lead to divergent classifications. For example, background (nominally non-wetland) points typically have relatively higher cartographic depth to water, so background points with lower DTW values are more frequently over-predicted by the model. However it must be remembered that divergent classification is relative to the MDW and some amount of divergence–especially over-prediction–is a fundamental goal of the analysis.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

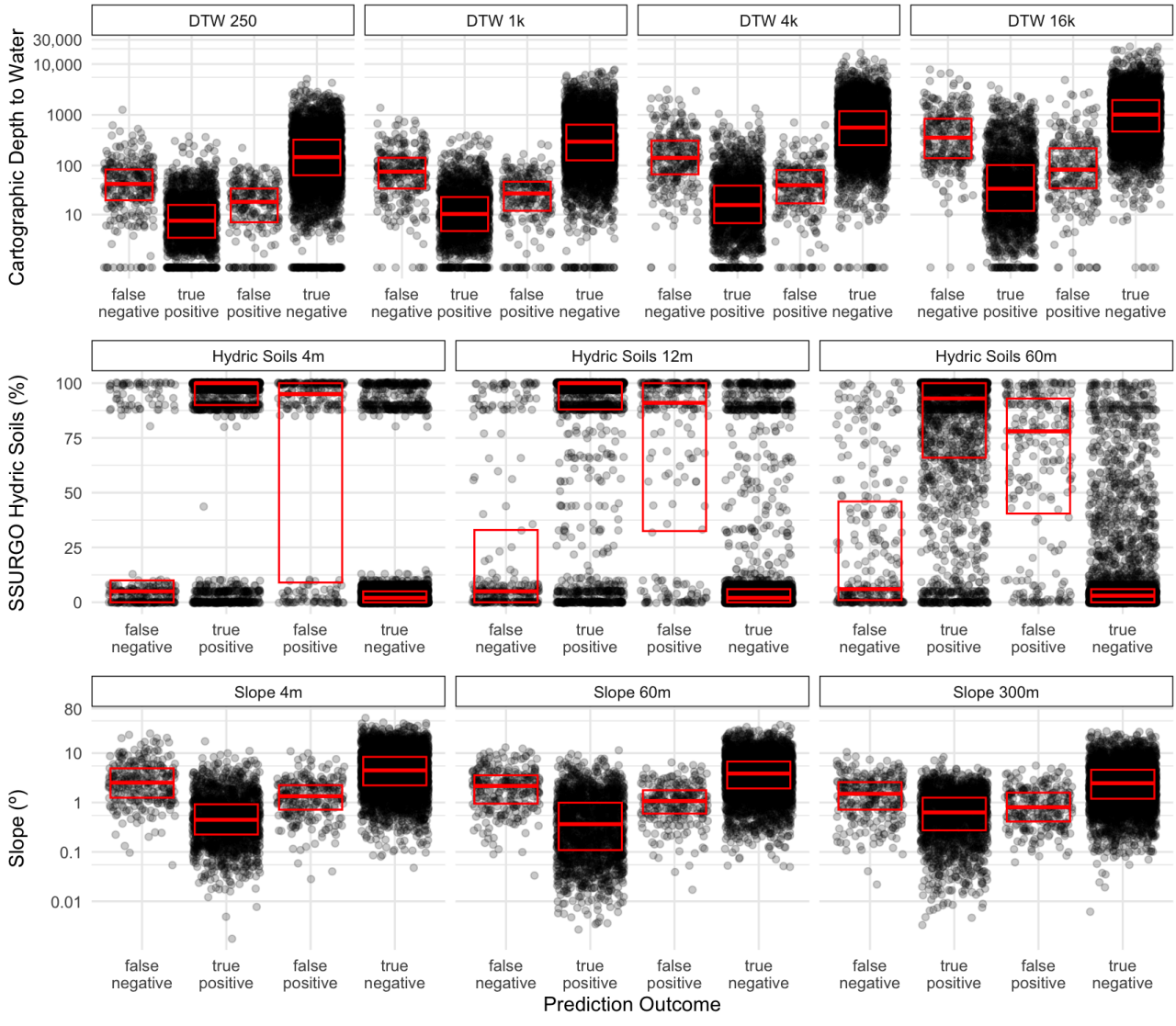

Figure 6. Distribution of predictor values across prediction outcomes for the three most important multi-scale predictors: cartographic depth to water, percent hydric soils, and slope. Prediction outcomes are based on binning the wetland probability score into a dichotomous (wetland/upland) outcome at a .5 threshold and comparing with the 2005 MassDEP Wetland layer. Red boxes show the median and interquartile range of predictor values. Results show that atypical values for important predictors lead to classifications that diverge from the reference layer.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

3.3 Model performance by wetland type and landcover

DEP wetland types (as given in the MDW) were crosswalked with land cover categories in the MassGIS 2016 Land Use/Land Cover layer to facilitate comparison with model predictions (Fig. 7). To supplement the land cover crosswalk, predicted wetlands that overlapped with coastal wetlands appearing in the Northeast Oceans dataset were assigned to the Salt Marsh category. The model excelled at identifying additional forested wetland areas (i.e. 'Wooded Swamp'), adding 167,329 acres of probable wetland, representing a 46% increase over the state inventory. Conversely, the model only identified 84% (57,518 acres) of the mapped forested wetland area appearing in the DEP inventory, suggesting that the even combined DEP and model results might seriously underestimate the extent of some cryptic wetlands. To existing estimates of freshwater marsh, model predictions added 96,443 acres, a 12% increase, and correctly identified 91% of DEP-designated freshwater marsh area. For each of the other crosswalked categories (excluding 'Other'), the model identified some but not all of the DEP-designated wetland area: Salt Marsh (94%), Cranberry Bog (93%), and Tidal Flat/Rocky Shore (46%). Results in the 'Other' category include land use designations *Impervious, Developed Open Space, Bare Land, Pasture/Hay,* and *Cultivated,* and likely encompass many historic wetland areas that have been lost to agricultural, residential, or urban development.

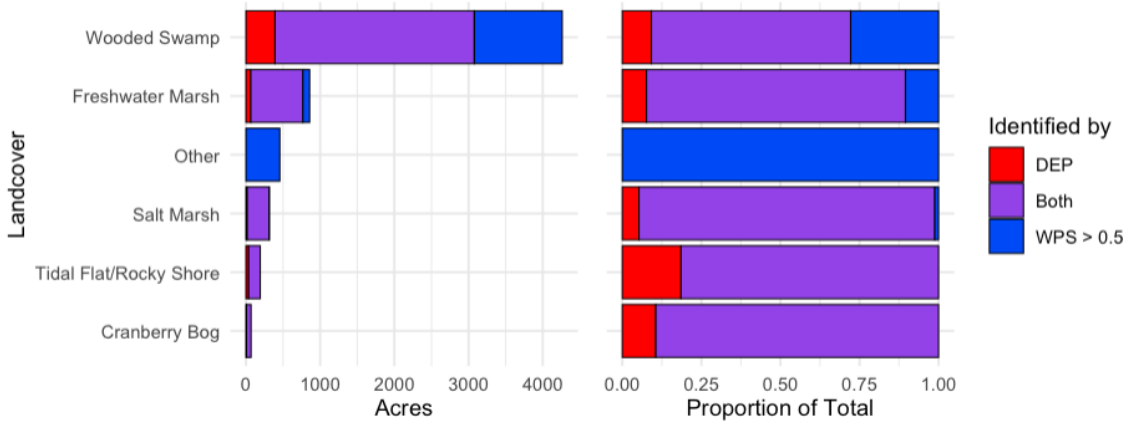

Figure 7. Comparison of DEP wetland layer and model predictions across wetland types. Comparison of raw Wetland Probability Score (dichotomized at 0.5) with DEP designation. Landcovers in the 'Other' category include, in descending order of acreage, Impervious, Developed Open Space, Bare Land, Pasture/Hay, Cultivated, Water, Palustrine Aquatic Bed, and Unconsolidated Shore.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

3.4 Impact on estimates of wetland extent

After post-processing to remove impervious surface and barren lands, the model identified an additional 225,781 acres of probable wetland compared to the MDW inventory–an increase of 40% (Table 4, ecoregions shown in Fig. 3). The Northeastern Highlands had the highest proportional increase at 49%, likely due to the dominance of the forested highland landscapes for which wetland detection via visual assessment methods is difficult. The Atlantic Coastal Pine Barrens, dominated by sandy soils, had the smallest absolute and proportional increase in wetland area. The 15% of additional wetland area here represents approximately 1/3rd of the proportional increase of the other two ecoregions.

Table 4. Additional wetland area identified across the ecoregions of Massachusetts

Ecoregion	2005 DEP	2023 Additional predicted (at .5 threshold)	Total	Percent increase
Atlantic Coastal Pine Barrens	78,402 ac	11,430 ac	89,832 ac	15%
Northeastern Coastal Zone	396,695 ac	171,085 ac	567,780 ac	43%
Northeastern Highlands	88,153 ac	43,266 ac	131,419 ac	49%
Total	563,250 ac	225,781 ac	789,031 ac	40%

4 Discussion

4.1 Overview

A random forest model was trained on the 2005 MassDEP Wetlands layer (MDW), using widely available open source software and publicly available data, and generated a Wetland Probability Score for the entire state of Massachusetts at a 4m resolution. The model has performed well by all metrics, and provides support for the use of machine learning to quickly and efficiently generate large-scale maps of wetland probabilities that can serve as an adjunct to conventional field and desk methods. The balance of over- and under-prediction relative to the MDW suggests that model is fit for the purpose of providing a more inclusive estimate of wetland extent in Massachusetts by modeling the extent of additional wetlands that are not captured in the MDW. Our results suggest a promising avenue for bootstrapping existing wetland inventories by using them as training data to identify additional probable wetland acres.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Adding 46% to existing estimates of forested wetlands in the state, results suggest that models such as this may be especially useful for identifying cryptic, previously unmapped forested wetlands. These wetlands are especially difficult to identify via the methods used by both NWI and the MDW. Efforts to combat climate change make it critical to understand the extent and location of forested wetlands, as they represent greater stores of aboveground carbon than other wetland types and greater stores of soil organic carbon than other forests. The model also predicted 69,761 acres of wetland in developed and agricultural areas (i.e. the 'Other' category of Fig. 7), entirely absent from the MDW, suggesting that this model may also be of use to understand long-term land use change, by estimating the extent and location of wetlands that have been lost to development of various kinds. The identification of topographically suitable areas with a high wetland probability may also be useful for the identification of potential sites for wetland restoration.

Examination of model performance and variable importance together provide support for the use of multi-scale terrain derivatives in wetland detection, and further expansion of multi-scalar approaches to other types of predictors. Variable importance scores show that model performance was dependent on the inclusion of terrain derivatives (e.g. cartographic depth to water and slope) at multiple scales, reinforcing the efficacy of multi-scale terrain derivatives established in previous work on wetland detection (Halabisky et al., 2023). More novel results include the positive performance of non-terrain multi-scale predictors, principally the SSURGO-derived Hydric Soils. The model was improved by inclusion of Hydric Soils at three different scales, and the larger-scale versions (smoothed at 12m and 60m resolutions) scored higher on importance than the baseline 4m scale. This finding provides support for the use of multi-scale smoothing with soil map data, both to account for neighborhood effects and to ameliorate the issues caused by crisp boundaries and their tendency to obscure underlying heterogeneity and varying transition zones.

4.2 Limitations

The use of the MDW for training and testing has implications for interpretation of model results and accuracy metrics: all accuracy metrics are relative to the reference dataset. Over-prediction, or false positives, should therefore be assumed to be a mix of false positives and real, cryptic wetlands that are missing from MDW. Likewise, under-prediction or false negatives are inevitably a mix of the model failing to identify an existing wetland, and false positives in the MDW where the model correctly identifies an upland. We expect the MDW itself is much more prone to false negatives than false positives, and that over-prediction by our model is therefore more informative than under-prediction.

Results show that the model failed to identify significant portions of DEP-designated wetlands: Tidal Flat/Rocky Shore (54%), Wooded Swamp (16%), Freshwater Marsh (9%), Cranberry Bog (7%), and Salt Marsh (6%). Researchers were also able to identify some patterns in over- and under-prediction through visual inspection of results and comparison with familiar sites and a small number of digitized field delineations. The model is prone to identifying low lying urban areas and

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

drained agricultural fields as wetlands. As noted above, these areas are topographically suited to be wetlands and prior to development they likely were. The model is also under-predicting certain wetland configurations, including sloped wetlands and small wetlands high in the drainage network. Future efforts to better capture these wetland types and/or environmental configurations could include oversampling these areas or fitting a weighted model.

4.3 Next Steps for Machine Learning and Wetland Mapping

The strong performance and relative speed of this approach suggest many useful avenues for further development. Releasing a publicly available GIS layer based on the Wetland Probability Score could be an aid to researchers as well as a useful screening tool for the town conservation commissions who, in Massachusetts, are charged with assessing the veracity of wetland delineations submitted by project proponents, as well as with desktop evaluation of potential wetlands requiring field delineation. Moving forward, this model or ones very similar to it are likely to be critical in large-scale assessments of soil carbon across the landscape–not only in wetlands (Stewart et al., 2024). It's also possible that a very similar model, with additional development, could be used to predict not only wetland extent but also wetland type.

In the domain of model performance and overall value, this project opens up several pathways for exploration and further improvement. First and foremost, the integration of field data into the training and/or validation process would almost certainly improve the performance of the model as well as reinforcing its validity for researchers, policymakers, and the public. One potential avenue for acquiring data at the necessary scale would be creating a centralized repository and unified format for field delineation data gathered for wetland permit applications. Conversely, a map of wetland probabilities could be used to inform a field sampling plan for ongoing carbon inventory and accounting (Stewart et al., 2024).

The development of this model was focused narrowly on the goal of estimating the extent of cryptic wetlands in aggregate (and by implication providing proof of concept for digital wetland mapping using machine learning), rather than producing the most accurate predictions on a pixel-by-pixel basis, or producing the fastest or most parsimonious model. As such, results suggest several low hanging fruits for further investigation and improvement of model efficiency and performance. For example, it is possible to generate variable importance metrics on a pixel-by-pixel basis in the training and/or testing data. These case-level metrics would support more detailed understanding of key topics such as which predictors support the identification of atypical wetlands currently underpredicted by the model, and potentially improving predictive performance with variable weighting and/or oversampling strategies. Additionally, there are several other ways to further refine the feature selection process, such as testing the importance of highly correlated predictors by fitting models with only the single most important predictor from each family, i.e. terrain, vegetation, soils, and so on. This could also be applied to a deeper investigation of the impact of multi-scale approach, by fitting models with only the single most important scale for each predictor and assessing model performance. In the interest of identifying the most lightweight and parsimonious model that

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

can achieve required levels of accuracy, these investigations should be accompanied by assessment of other strategies for tuning model hyperparameters such as number of trees, size of bootstrap sample, and number of variables selected for each decision tree.

5 Conclusion

The status of wetlands is a critical issue that is only becoming more important as the climate crisis intensifies. Monitoring, managing, and protecting these vital ecosystems, safeguarding the many functions they perform, all depend on a strong understanding of their extent and location. This is a critical issue for cryptic wetlands that are missing from most inventories, often due to forest land cover, which represent especially dense pools of organic carbon. Fast and efficient processes for detecting probable wetlands, using free, open source software and publicly available data, are a needed adjunct to conventional methods for wetland mapping and monitoring.

Competing interests: The authors declare that they have no conflict of interest.

Author contributions. RSF and SG designed the data pre-processing and sampling protocol. SG and EG conceptualized and carried out all operations in QGIS: data acquisition, pre-processing, sampling, and post-processing. RF designed and carried out all operations in R, including data wrangling and model training, testing, and prediction processes, and data visualizations. KZD provided project administration and support. RSF prepared the paper with contributions from all authors.

Acknowledgments. The authors would like to acknowledge Jenny Watts, Gillian Davies, and Matthew von Walde for their input and support early in the development of this project. Initial development of the wetland probability map was conducted as part of a project for the Massachusetts Department of Environmental Protection. Contents of this article are solely the responsibility of the authors; the views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of the government of Massachusetts or any of its representatives.

Data and code availability: All datasets used in this analysis are publicly available. R code is available on request.

References

Northeast Ocean Data Portal: https://www.northeastoceandata.org/, last access: 1 May 2024.

Protecting Wetlands in Massachusetts | Mass.gov: https://www.mass.gov/info-details/protecting-wetlands-in-massachusetts, last access: 7 March 2024.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Baker, C. D., Polito, K. E., Beaton, M. A., and Suuberg, M.: Inland and Coastal Wetlands of Massachusetts: Status and Trends, Commonwealth of Massachusetts, 2019.

Behrens, T., Viscarra Rossel, R. A., Kerry, R., MacMillan, R., Schmidt, K., Lee, J., Scholten, T., and Zhu, A.-X.: The relevant range of scales for multi-scale contextual spatial modelling, Sci. Rep., 9, 14800, https://doi.org/10.1038/s41598-019-51395-3, 2019.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm. Remote Sens., 114, 24–31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Davidson, N. C.: How much wetland has the world lost? Long-term and recent trends in global wetland area, Mar. Freshw. Res., 65, 934, https://doi.org/10.1071/MF14173, 2014.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognit. Lett., 27, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Felton, B. R., O'Neil, G. L., Robertson, M.-M., Fitch, G. M., and Goodall, J. L.: Using Random Forest Classification and Nationally Available Geospatial Data to Screen for Wetlands over Large Geographic Regions, Water, 11, 1158, https://doi.org/10.3390/w11061158, 2019.

Fluet-Chouinard, E., Stocker, B. D., Zhang, Z., Malhotra, A., Melton, J. R., Poulter, B., Kaplan, J. O., Goldewijk, K. K., Siebert, S., Minayeva, T., Hugelius, G., Joosten, H., Barthelmes, A., Prigent, C., Aires, F., Hoyt, A. M., Davidson, N., Finlayson, C. M., Lehner, B., Jackson, R. B., and McIntyre, P. B.: Extensive global wetland loss over the past three centuries, Nature, 614, 281–286, https://doi.org/10.1038/s41586-022-05572-6, 2023.

Gale, S.: Automated Identification of Wetlands Using GIS in North Carolina, North Carolina Dept. of Environmental Quality, Div. of Water Resources, 2021.

Griffith, G. E., Omernik, J. M., Bryce, S. A., Royte, J., Hoar, W. D., Homer, J. W., Keirstead, D., Metzler, K. J., and Hellyer, G.: Ecoregions of New England (2 sided color poster with map, descriptive text, summary tables, and photographs), Rest. VA US Geol. Surv. Scale 11325000, 2009.

Halabisky, M., Miller, D., Stewart, A. J., Lorigan, D., Brasel, T., and Moskal, L. M.: The Wetland Intrinsic Potential tool: Mapping wetland intrinsic potential through machine learning of multi-scale remote sensing proxies of wetland indicators, EGUsphere, 1–19, https://doi.org/10.5194/egusphere-2022-665, 2022.

Halabisky, M., Miller, D., Stewart, A. J., Yahnke, A., Lorigan, D., Brasel, T., and Moskal, L. M.: The Wetland Intrinsic Potential tool: mapping wetland intrinsic potential through machine learning of multi-scale remote sensing proxies of wetland indicators, Hydrol. Earth Syst. Sci., 27, 3687–3699, https://doi.org/10.5194/hess-27-3687-2023, 2023.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X.: Applied Logistic Regression, John Wiley & Sons, 528 pp., 2013.

Hunter, G. J., Bregt, A. K., Heuvelink, G. B. M., De Bruin, S., and Virrantaus, K.: Spatial Data Quality: Problems and Prospects, in: Research Trends in Geographic Information Science, edited by: Navratil, G., Springer Berlin Heidelberg, Berlin, Heidelberg, 101–121, https://doi.org/10.1007/978-3-540-88244-2_8, 2009.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

Ishwaran, H. and Kogalur, U. B.: Random survival forests for R, R News, 7, 25–31, 2007.

Jafarzadeh, H., Mahdianpari, M., Gill, E. W., Brisco, B., and Mohammadimanesh, F.: Remote Sensing and Machine Learning Tools to Support Wetland Monitoring: A Meta-Analysis of Three Decades of Research, Remote Sens., 14, 6104, https://doi.org/10.3390/rs14236104, 2022.

Kudray, G. M. and Gale, M. R.: Evaluation of National Wetland Inventory maps in a heavily forested region in the upper Great Lakes, Wetlands, 20, 581–587, https://doi.org/10.1672/0277-5212(2000)020[0581:EONWIM]2.0.CO;2, 2000.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, 2002.

Martin, G. I., Kirkman, L. K., and Hepinstall-Cymerman, J.: Mapping Geographically Isolated Wetlands in the Dougherty Plain, Georgia, USA, Wetlands, 32, 149–160, https://doi.org/10.1007/s13157-011-0263-7, 2012.

Matthews, J. W., Skultety, D., Zercher, B., Ward, M. P., and Benson, T. J.: Field Verification of Original and Updated National Wetlands Inventory Maps in three Metropolitan Areas in Illinois, USA, Wetlands, 36, 1155–1165, https://doi.org/10.1007/s13157-016-0836-6, 2016.

Maxwell, A. E., Warner, T. A., and Strager, M. P.: Predicting Palustrine Wetland Probability Using Random Forest Machine Learning and Digital Elevation Data-Derived Terrain Variables, Photogramm. Eng. Remote Sens., 82, 437–447, https://doi.org/10.14358/PERS.82.6.437, 2016.

Miller, B. A., Koszinski, S., Wehrhan, M., and Sommer, M.: Impact of multi-scale predictor selection for modeling soil properties, Geoderma, 239–240, 97–106, https://doi.org/10.1016/j.geoderma.2014.09.018, 2015.

Mitsch, W. J. and Gosselink, J. G.: Wetlands, John Wiley & Sons, 744 pp., 2015.

Murphy, P., Ogilvie, J., and Arp, P.: Topographic modelling of soil moisture conditions: A comparison and verification of two models, Eur. J. Soil Sci., 60, 94–109, https://doi.org/10.1111/j.1365-2389.2008.01094.x, 2009.

Murphy, P. N. C., Ogilvie, J., Meng, F.-R., White, B., Bhatti, J. S., and Arp, P. A.: Modelling and mapping topographic variations in forest soils at high resolution: A case study, Ecol. Model., 222, 2314–2332, https://doi.org/10.1016/j.ecolmodel.2011.01.003, 2011.

Nahlik, A. M. and Fennessy, M. S.: Carbon storage in US wetlands, Nat. Commun., 7, 1–9, https://doi.org/10.1038/ncomms13835, 2016.

Nikiforova, A. A., Fleis, M. E., Nyrtsov, M. V., Kazantsev, N. N., Kim, K. V., Belyonova, N. K., and Kim, J. K.: Problems of modern soil mapping and ways to solve them, CATENA, 195, 104885, https://doi.org/10.1016/j.catena.2020.104885, 2020.

The Wetland Identification Model (WIM) – A New Arc Hydro Functionality for Predicting Wetland Locations Using LiDAR Elevation Data and Machine Learning: https://community.esri.com/t5/water-resources-blog/the-wetland-identification-model-wim-a-new-arc/ba-p/884298, last access: 23 January 2024.

O'Neil, G. L., Goodall, J. L., and Watson, L. T.: Evaluating the potential for site-specific modification of LiDAR DEM derivatives to improve environmental planning-scale wetland identification using Random Forest classification, J. Hydrol., 559, 192–208, https://doi.org/10.1016/j.jhydrol.2018.02.009, 2018.

https://doi.org/10.5194/egusphere-2024-3326
Preprint. Discussion started: 2 January 2025

QGIS Development Team: QGIS Geographic Information System, 2023.

R Core Team: R: A language and environment for statistical computing, 2023.

Rapinel, S., Panhelleux, L., Gayet, G., Vanacker, R., Lemercier, B., Laroche, B., Chambaud, F., Guelmami, A., and Hubert-Moy, L.: National wetland mapping using remote-sensing-derived environmental variables, archive field data, and artificial intelligence, Heliyon, 9, e13482, https://doi.org/10.1016/j.heliyon.2023.e13482, 2023.

Soil Survey Staff: SSURGO Metadata - Table Column Descriptions Report, Natural Resources Conservation Service, United States Department of Agriculture, Washington, D.C, 2015.

Soil Survey Staff: Soil Survey Geographic (SSURGO) Database for Massachusetts, n.d.

Stewart, A. J., Halabisky, M., Babcock, C., Butman, D. E., D'Amore, D. V., and Moskal, L. M.: Revealing the hidden carbon in forested wetland soils, Nat. Commun., 15, 726, https://doi.org/10.1038/s41467-024-44888-x, 2024.

Stolt, M. H. and Baker, J. C.: Evaluation of national wetland inventory maps to inventory wetlands in the southern Blue Ridge of Virginia, Wetlands, 15, 346–353, https://doi.org/10.1007/BF03160889, 1995.

Stone, J. R., Stone, B. D., DiGiacomo-Cohen, M. L., and Mabee, S. B.: Surficial materials of Massachusetts—A 1:24,000-scale geologic map database, Scientific Investigations Map, U.S. Geological Survey, https://doi.org/10.3133/sim3402, 2018.

Tiner, R. W.: Use of high-altitude aerial photography for inventorying forested wetlands in the United States, For. Ecol. Manag., 33–34, 593–604, https://doi.org/10.1016/0378-1127(90)90221-V, 1990.

Ting, K. M.: Confusion Matrix, in: Encyclopedia of Machine Learning, edited by: Sammut, C. and Webb, G. I., Springer US, Boston, MA, 209–209, https://doi.org/10.1007/978-0-387-30164-8_157, 2010.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G.: Modelling species presence-only data with random forests, Ecography, 44, 1731–1742, https://doi.org/10.1111/ecog.05615, 2021.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J.: Predictive performance of presence-only species distribution models: a benchmark study with reproducible code, Ecol. Monogr., 92, e01486, https://doi.org/10.1002/ecm.1486, 2022.