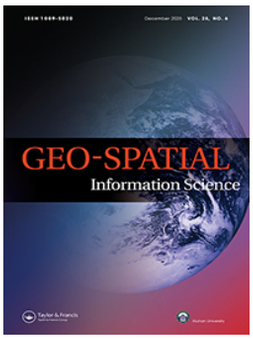

Geo-spatial Information Science

ISSN: 1009-5020 (Print) 1993-5153 (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

Understanding forest insect outbreak dynamics: a comparative analysis of machine learning techniques

Roberto Molowny-Horas, Saeed Harati-Asl & Liliana Perez

To cite this article: Roberto Molowny-Horas, Saeed Harati-Asl & Liliana Perez (22 Jul 2025): Understanding forest insect outbreak dynamics: a comparative analysis of machine learning techniques, Geo-spatial Information Science, DOI: 10.1080/10095020.2025.2529992

To link to this article: https://doi.org/10.1080/10095020.2025.2529992

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2025.2529992

Taylor & Francis
Taylor & Francis Group

Understanding forest insect outbreak dynamics: a comparative analysis of machine learning techniques

Roberto Molowny-Horas [a], Saeed Harati-Asl [b] and Liliana Perez [c]

[a]The Ecosystem Modelling Facility, CREAF, Cerdanyola del Vallès, Spain; [b]Roger Tomlinson laboratory, Department of Geography, McGill University, Montreal, Canada; [c]Laboratoire de Géosimulation Environnementale (LEDGE), Département de Géographie, Université de Montréal, Montréal, Canada

ABSTRACT

Accurate modeling and simulation of forest land cover change resulting from epidemic insect outbreaks play a crucial role in equipping scientists and forest managers with essential insights. These insights enable proactive planning and the formulation of effective strategies to mitigate the impact of such disturbances. By employing advanced modeling techniques, researchers and managers can anticipate the evolving dynamics of forest ecosystems, thereby facilitating timely interventions and sustainable management practices. In this study, we applied sixteen machine-learning models, plus two ensemble averaging procedures, to Mountain Pine Beetle (*Dendroctonus ponderosae*) infestation data in British Columbia, to calculate projections of insect-induced deforestation. Model drivers included topographic, climatic and adjacency variables. We verified the results of the simulations by randomly splitting datasets between training and test subsets (aka Validation assessment), as well as by comparing future projections with observations (aka Prediction assessment). All calculations were carried out for different mountain pine beetle map sets and time differences, and we employed up to seven performance metrics (six threshold-dependent and one threshold-independent) and four error metrics to assess goodness of prediction. ANCOVA tests were then run on metric results to test differences between Validation and Prediction assessments. In addition, we computed Friedman rankings for all simulation and metrics. Our results showed that validation assessments were, most of the time, significantly more optimistic than prediction assessments. We also noted that different conclusions could be reached for different performance metrics. We conclude that, for prediction purposes, error metrics and components of the confusion table were most helpful in understanding the ability and limitations of Mountain Pine Beetle predictive maps. These results also suggest that, in general, care must be taken in assessing prediction performance of machine-learning models based solely on validation tests.

ARTICLE HISTORY
Received 27 March 2024
Accepted 2 July 2025

KEYWORDS
Land cover changes; forest disturbances; model calibration; machine-learning; cross-validation; map comparison; insect outbreaks; *Dendroctonus ponderosae*

1. Introduction

Disturbances are a critical component of forests dynamics, which shape and substantially affect these key ecosystems (Bourbonnais, Nelson, and Wulder 2014; McCullough, Werner, and Neumann 1998). Particularly of interest, forest insect epidemics could exert a severe impact on ecosystem dynamics due to enhanced mortality or growth reduction of millions of trees over widespread areas (Axelson, Alfaro, and Hawkes 2010; Kastridis et al. 2022; MacLean 2016; Pelz and Smith 2012; Robbins 2008). In the province of British Columbia (BC), Canada, an unprecedented insect outbreak of mountain pine beetle (*Dendroctonus ponderosae* Hopkins; hereafter MPB), feeding mainly on lodgepole pine (Pinus contorta), started in the early 90s and reached a peak between 2005 and 2006, which facilitated a massive migration of beetles into the province of Alberta (Patriquin, Wellstead, and White 2007; Petersen and Stuart 2014; Strohm, Reid, and Tyson 2016). Only in British Columbia, the total cumulative loss of marketable pine timber, due to the MPB epidemic, was estimated at 752 million cubic meters (58% of sellable pine volume in the province) by 2017, corresponding approximately to more than 16 million of the 55 million hectares of forest in BC (Bleiker 2019; Corbett et al. 2016; Government of British Columbia 2023). It is thus a pressing matter to be able to anticipate the extent and severity of such disturbances in the short and middle term, especially with the spread of MPB infestations beyond BC and into the Canadian boreal forest. To build models that simulate the spread of MPB infestations, we must choose those mathematical methodologies that are flexible enough to adapt to the available datasets.

Machine-learning (hereafter, ML) techniques are a set of methodologies within the field of statistical learning whose objectives are to understand the structure of the datasets, to uncover the underlying trends

CONTACT Liliana Perez l.perez@umontreal.ca
Supplemental data for this article can be accessed online at https://doi.org/10.1080/10095020.2025.2529992.

or patterns without a priori information and to produce models in order to make predictions (Alpaydın 2014). Their inherent flexibility makes them an optimum choice for classification or regression of complex problems by learning from the data without imposing any strict constraint on the relationship between predictors and response variables (James et al. 2013; Mahdizadeh Gharakhanlou and Perez 2024; Patel et al. 2021; Zhao et al. 2025). That is, ML techniques excel at identifying patterns from the data without any extra input from the researcher. In the particular case of the supervised ML classification algorithms that we will be implementing in the present study, only an identification of the response variable is required.

The assessment of the predictive capabilities of ML classification techniques has become a central issue in this field. A correct determination of their performance is thus critical, since we aim to apply those ML methodologies to actual cases for which useful and accurate answers are needed, and from which decisions must be drawn. There are quite a few performance indices available for categorical data, most of them derived from combinations of elements in the confusion matrix. However, past research has shown how hard it is to summarize the predictive capabilities of an ML algorithm into a single metric (Pontius 2022).

An additional problem in ML classification lies in the fact that performance assessment may be hindered or even biased when the response variable shows noticeable class imbalance, i.e. low prevalence of one of the thematic classes (He and Garcia 2009). Metrics derived from the confusion matrix may then over or underestimate systematically the performance of the ML algorithm. For example, if we predicted a 100% failure for a binary dataset which in fact has a success probability of 1% (thus, a prevalence of 0.01) the resulting Accuracy index would be 0.99, thus indicating a very successful prediction, even though the model did not predict any of the cases of success at all. Inconsistencies like this must be handled correctly lest we end up with an ineffective model.

Finally, an extra difficulty arises when one intends to implement an ML model to elaborate future predictions of a natural phenomenon that varies in time and space, and whose spatial spread depends partly on the state of neighboring locations, like the MPB outbreak described above. As a consequence, the conditions under which the models are initially calibrated may change after some time and, thus, the assessment of their predictive capabilities may become even more challenging.

Our main research objective is to carry out a comprehensive evaluation of the performance of different ML classification models on MPB binary disturbance data when those models are employed to track the spread of the MPB infestation in the near future. Due to the particularities of the available MPB maps (e.g. varying spatial and temporal dynamics, low prevalence), we seek to determine which performance metrics are or are not relevant for our low-prevalence, spatially and temporally dynamic datasets. We thus set out to answer the following questions:

(1) How do different ML models compare when they are used to calculate the extent of a spatially-spreading disturbance in the near future?
(2) How can we best assess their performance?

To answer these questions, we applied various metrics to compare predictions with reference data. We then noted whether metrics showed differences among predictions, and explored the reasons for such differences or lack thereof.

2. Materials and methods

A thorough depiction of the study area and the data sets used in this work have already been described in previous publications (Harati, Perez, and Molowny-Horas 2020), and thus only a brief description will be provided below.

2.1. Study area and pine mortality data

We conducted our study with data selected from a section of the Canadian province of British Columbia (BC) covering an area of 944,735 km^2. BC is characterized by a very diverse mountainous landscape that is subject to a variety of disturbance regimes (Axelson, Alfaro, and Hawkes 2009; Haughian et al. 2012; Klenner et al. 2008). Climate in the region is driven by the vicinity of the Pacific Ocean to the west, the influence of continental air masses in the plateaus and the presence of the Rocky Mountains to the east (Lemmen et al. 2008). We used the "kgc" R package (Bryant et al. 2017) to determine the Köppen-Geiger climate classification of the forested area. The preponderant climate class was Dfc (subarctic), although there were also Dfb (warm-summer humid continental), ET (tundra), Dsc (dry-summer subarctic), Dsb (warm-summer Mediterranean continental) and Cfb (oceanic) zones, in decreasing order of spatial extension. BC is considered the most physically and biologically diverse region in Canada, thanks to its proximity to the Pacific Ocean and the presence of numerous mountain chains. These mountains divide the province creating a series of valleys and a broad central interior plateau; two major ranges are the Coast Mountains, which lie in the western part of the province, and the Canadian portion of the Rocky Mountains in the eastern part. Within and between these zones, biodiversity varies, while local disturbances, such as fire, insects, disease, windthrow and human activity, significantly influence species

distributions (McGillivray 2011). Finally, about 70% of the study area (Figure 1) is covered by forests, mostly of the lodgepole (Pinus contorta) pine species.

Spatial observations of lodgepole pine mortality due to MPB attacks were collected and processed regularly by the BC Ministry of Forests from Aerial Overview Surveys and LANDSAT satellite images between 1999 and 2014 (BC Ministry of Forests 2015), and made available at their web site. The resulting raster maps had a spatial resolution of 400 × 400 m per pixel and specified the proportion of each pixel that showed evidence of pine mortality due to MPB attacks. After downloading those maps, we converted their relative proportion scale, which ranged from 0 to 1000, into a binary scale indicating presence (1) or absence (0) of MPB infestation. Those binary maps were calculated by applying a threshold to the original proportion data. For details, see (Harati, Perez, and Molowny-Horas 2020).

2.2. Explanatory variables

To account for past dynamics of the infestation adjacent to each map pixel, we also included the infestation status from the previous year as explanatory variable. In short, let us denote by t_1 the starting date for the simulation, and by t_2 the future year for which a prediction is required. Then, the year previous to t_1 will be symbolized by $t_{1p} = t_{1-1}$, such that $t_{1p} < t_1 < t_2$.

Table 1 shows a list of other topographic, climatic and adjacency explanatory variables that we selected based on the available literature (Cooke and Carroll 2017; Raffa et al. 2008; Safranyik and Carroll 2007) and expert knowledge about the phenology and dynamics of the lodgepole pine and the MPB species. They thus represented the most important drivers with which to feed our ML models. The DEM model was downloaded from the Open Maps collection in Open Government Portal (Government of Canada 2024). Previous work highlighted the importance of distance variables that describe neighborhood effects on the spread of MPB. Further details can be found in (Harati, Perez, and Molowny-Horas 2020). As a side note, it must be mentioned that we did not include in Table 1 variables that accounted for any likely dependence on meteorology (e.g. hot or cold weather, severe drought or heavy rain, windy condition), due to the difficulty in collecting those data.

2.3. Machine-learning modeling and ensemble learnings

2.3.1. Software choice

Our goal was to model and simulate the presence of MPB infestation in the lodgepole pine forest cover within the province of BC. To that aim, we turned to the "caret" R package (Kuhn et al. 2023) and the numerous ML methodologies available therein. As

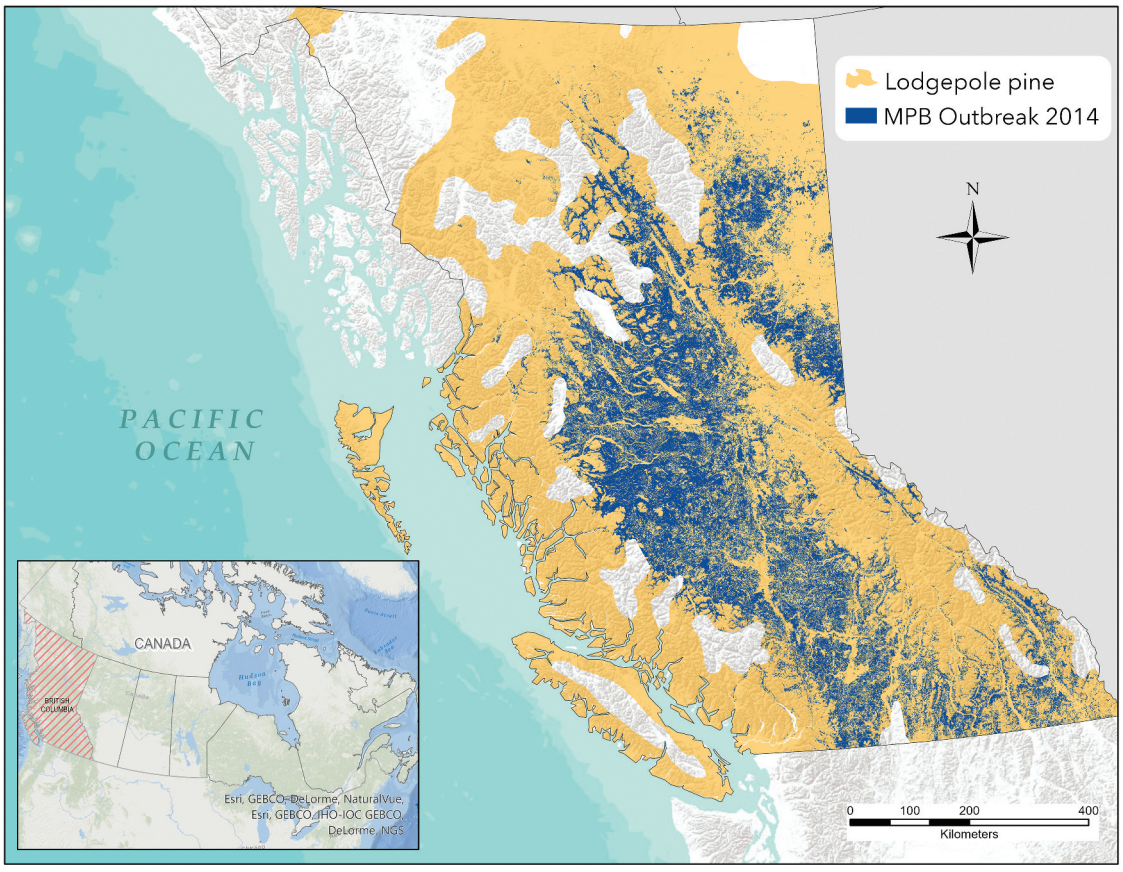

Figure 1. Map of the study area, lodgepole pine forest cover and MPB infested areas of the province of British Columbia.

Table 1. List of predictor variables used for modeling. The "time" column indicates whether the variable corresponds to t_{1p} or t_1 (see text for an explanation). The "range" column specifies the range of values across the whole study area before normalization.

Predictor description	Units	Time	Range
Elevation	m	–	0–5433
Sine aspect	Radians	–	−1–1
Cosine aspect	Radians	–	−1–1
Ruggedness	Arbitrary	–	−693–4186
Max. temp. warmest month	°C	–	2.5–29.6
Mean temp. coldest quarter	°C	–	−21.6–6.5
Annual precipitation	mm	–	251–4607
Identity	Arbitrary	t_{1p}	–
Linear weight	Arbitrary	t_{1p}	–
Inverse-distance weight	Arbitrary	t_{1p}	–
Square-inverse-distance weight	Arbitrary	t_{1p}	–
Identity	Arbitrary	t_1	–
Linear weight	Arbitrary	t_1	–
Inverse-distance weight	Arbitrary	t_1	–
Square-inverse-distance weight	Arbitrary	t_1	–

MPB data, we used the 16-year dataset on recorded MPB epidemics described above. Rather than implementing ML simulations using all the available methodologies, we used the ranking shown in (Fernández-Delgado et al. 2014) for binary classification methodologies. Then, we selected those models that had performed best in their tests and whose implementation was available in the aforementioned "caret" R package (see (Fernández-Delgado et al. 2014), Table 9, all algorithms with suffix "_t"). That way, we could remove algorithms that, a priori, might perform poorly in our tests. In all, we ended up with 16 ML independent classifiers (see Table 2).

Once the different ML algorithms have been trained on the same data one by one, we may obtain better predictive performance by combining the outputs of several ML algorithms into one single learner. These so-called ensemble methods seek to improve the predictions of different ML models by combining their individual outputs into a new set of predictions. The idea behind ensemble methods is that weaknesses and biases from each individual model can be mitigated, and their strengths enhanced, by choosing an optimum combination of models. In this work, we tried two different ensemble methods: a) a simple average (i.e. an unweighted sum of the probabilities) and b) stacking of the ML models with the "caretEnsemble" R package (Deane-Mayer and Knowles 2023). For the latter, we chose a greedy optimization on Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) plot via a generalized linear model (see "caretEnsemble" package for details). In all, we calculated 18 models (i.e. 16 ML models plus 2 ensemble models).

2.3.2. Global parameters

Several global parameters common to all simulations and models must be specified prior to the calculations, namely:

- Time difference between MPB maps, $t_2 - t_1$: larger time differences between MPB maps could show more changes (i.e. more infestations), increasing prevalence, at the price of reducing causality (i.e. infested pixels at t_1 are less likely to be the source of infestation of pixels at t_2 when $t_2 - t_1$ is large). We thus tried several values, namely $t_2 - t_1 = 1$, 2, 3, and 4 years. As specified, $t_{1p} = t_1 - 1$ always.

Table 2. List of the ML classifiers that have been selected for this study. The names in the leftmost column correspond to the names used in the "caret" package.

ML classifier	Methodology	R package	Package version
avNNet	Multilayer perceptrons	caret	6.0–86
bayesglm	Bayesian GLM	arm	1.11.2
cforest	Random forests	party	1.3.5
C50	Decision trees	C50	0.1.3.1
fda	Flexible discriminant	mda	0.5.2
knn	Nearest neighbors	caret	6.0–86
mlp	Multilayer perceptrons	RSNNS	0.4.12
mlpWeightDecay	Multilayer perceptrons	RSNNS	0.4.12
nnet	Neural networks	nnet	7.3.14
parRF	Random forests	randomForest	4.6.14
pcaNNet	Multilayer perceptrons	RSNNS	0.4.12
pda	Penalized discriminant	mda	0.5.2
rf	Random forests	randomForest	4.6.14
svmPoly	Support vector machines	kernlab	0.9.29
svmRadial	Support vector machines	kernlab	0.9.29
svmRadialCost	Support vector machines	kernlab	0.9.29

- Half-width of neighborhood window, d_{max}: a large window would span larger distances from previously infested pixels, further extending the radius of influence of the neighborhood-dependent predictors.
- Number of subsets for model tuning: tuning of hyper-parameters was carried out by randomly splitting the 10,000 points of the calibration dataset into 10 subset of 1000 points each (see description below).
- Number of random simulations per ML algorithm: we repeated the calibration + validation + prediction steps 5 times per every ML algorithm.

2.3.3. Model calibration, validation and prediction evaluation

We evaluated the performance of each ML model in three steps: a) calibration, b) validation and c) prediction assessments. Due to the considerable computer resources that were needed to run all simulations for the entire MPB maps, we could not use all pixels from the infestation maps. Instead, we ran steps a), b) and c) above on data subsets drawn randomly from each MPB map as follows:

(a) Calibration: first, from a set of three MPB maps at t_{1p}, t_1 and t_2 we drew independent and geographically coincident subsets of 10,000 points each. We then performed a random oversampling of points to finally have 5,000 points of new infestations and 5,000 points of persistence of non-infestation (for details see Section 2.1.3. Data normalization and imbalance correction). These points were then used to calibrate each ML algorithm.

(b) Validation: next, new, independent subsets of 10,000 points were drawn randomly from the same t_{1p}, t_1 and t_2 MPB maps, which was then used to test the performance of the model under the same conditions as the calibration phase.

(c) Prediction: finally, the prediction evaluation step involved randomly selecting subsets of 10,000 points from an independent dataset of future MPB infestations to evaluate whether or not the model could predict future infestation correctly. If we denote the time difference as $\Delta = t_2 - t_1$, the new dataset was drawn from MPB maps at t'_{1p}, t'_1 and t'_2 such that $t'_{1p} = t_{1p} + \Delta$, $t'_1 = t_1 + \Delta$ and $t'_2 = t_2 + \Delta$. Subsequently, the model used those subsets from t'_{1p} and t'_1 to predict and compare the observed infestation at t'_2. For example, in one simulation we set $\Delta = 4$ years and calibrated an ML model for $t_{1p} = 2005$, $t_1 = 2006$, $t_2 = 2010$, then we assessed its validation performance with new data from the same $t_2 = 2010$, but evaluated its prediction performance at $t'_2 = 2014$, with $t'_{1p} = 2009$ and $t'_1 = 2010$. Tables A1, A2, A3 and A4 in Appendix A of the Supplementary Material show the different combinations of t_{1p}, t_1 and t_2 that were used throughout the simulations. In all, we calculated $(13 + 11 + 9 + 7) \times 5 \times 2 \times (16 + 2) = 7200$ models, which correspond to the number of rows in those tables (13, 11, 9, 7), times the number of random repetitions (5 rounds of calculations to compensate for computational limits of working with large sample sizes), times the number of subsets (2 subsets for calibration and validation), times the number of models (16 ML models plus 2 ensemble models).

Model calibration comprised the tuning and training of the algorithm. Each ML model was calibrated independently with the aid of the built-in "train" function of the "caret" package. Tuning of an ML model is a preliminary step in which the best set of so-called hyperparameters is calculated. Those hyperparameters are a small set of internal parameters that control the learning/training process of every individual ML algorithm (see the documentation to the "caret" R package for further details). Thus, we applied a 10-fold cross-validation to split each of the 10,000-point calibration subset into 10 equally sized random subsets of 1,000 points each. The ML model was computed 10 times by holding out one of the folds each time. The best set of hyperparameters was the one that corresponded to the model with the best accuracy performance.

Subsequently, the model was trained by applying the ML algorithm to all 10,000 points of the Calibration data with the selected set of hyperparameters. All these processes took place automatically inside the "train" function. Next, we selected the optimal model by maximizing the Accuracy metric within "train".

Later, ML model validation was performed by calculating several performance and error metrics (see below). This step of the analysis was carried out by applying the trained model on the validation subset.

Finally, prediction evaluation included applying the trained model on the prediction subsets, drawn from MPB maps at t'_{1p}, t'_1 and t'_2, to assess how the model predicts independent future infestation scenarios under different conditions. The same performance and error metrics as before were computed. Figure 2 2 illustrates the different steps involved.

We used the models determined for every choice of t_{1p}, t_1, t_2 time points and calculated and evaluated the predicted infested pixels for $t_2 + \Delta$ time points, where Δ =1, 2, 3, and 4. Therefore, the conditions under which observations at $t_2 + \Delta$ had been generated

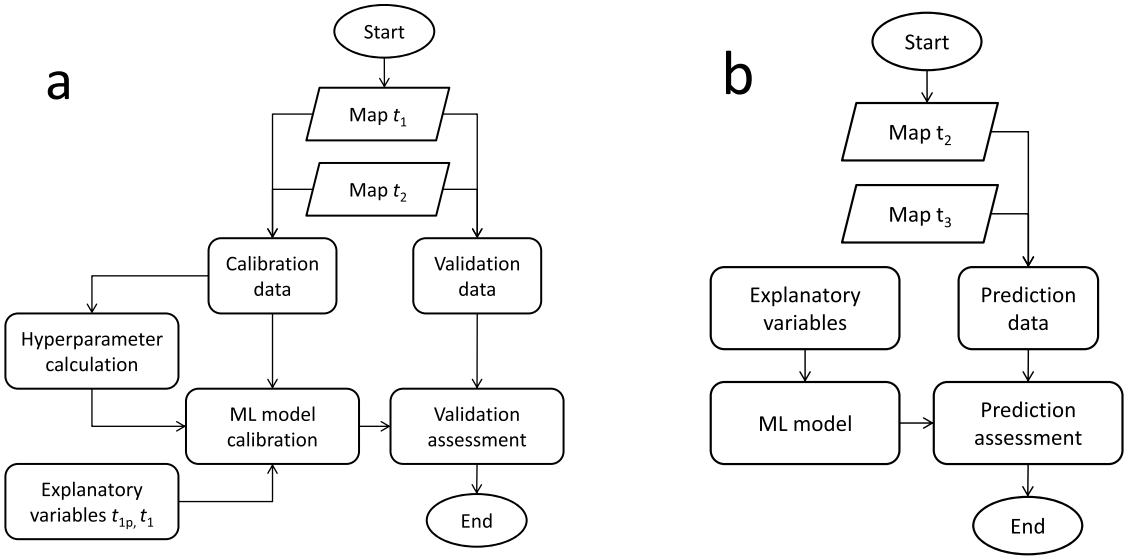

Figure 2. Flowchart of the a) calibration and validation, and b) prediction assessment stages. For the sake of clarity of the diagram the input adjacency map at t_{1p} has been omitted, since it can be considered as another explanatory variable. The "ML model" box in b) corresponds to the output of the "ML model calibration" of stage a). Time point $t_2 = t_1 + \Delta$ and $t_3 = t_2 + \Delta$, where $\Delta = 1, 2, 3$ or 4 (see text for details).

were new to the model. No single metric can probably give a full account of the performance of an ML classifier. Therefore, we calculated seven performance metrics, namely Accuracy, Specificity, Sensitivity, Precision, Cohen' κ, F1-score and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) plot. We also calculated four error metrics, namely False Positives, False Negatives, Quantity Error and Allocation Error. We describe in detail the characteristics of the above performance and error metrics in Appendix B of the accompanying Supplementary Material.

To provide a more comprehensive evaluation of ML models, incorporating qualitative results alongside the aforementioned quantitative metrics can offer additional insights. Qualitative visualizations, such as Venn diagrams, can effectively illustrate the overlap or distinction between true positives, false positives and false negatives across multiple models. This approach complements numerical metrics, enabling a more nuanced understanding of model performance and decision boundaries that are not easily captured by quantitative measures alone (Ho et al. 2021; O'Brien and Zhou 2018).

2.3.4. Data normalization and imbalance correction

The performance of ML models can be negatively affected by highly differing value ranges in the predictor dataset. To mitigate those effects, we normalized the topographic and climatic predictors to a [0, 1] scale. That performance can also be seriously affected by data imbalance (i.e. over or under-representation of one or several classes in the dataset; see e.g (He and Garcia 2009). To tackle imbalance problems, we randomly oversampled (without replacement) the infestation class (i.e. the 1's) such that there was the same number of 0's and 1's in the calibration subset (i.e. 5,000 points each). However, in the validation and prediction subsets we kept the original imbalance, since they should reflect the observed class distribution. Figure 3 summarizes the process of selection of data points for calibration, validation and prediction.

3. Results

3.1. Model validation and prediction evaluation

Top and bottom panels in Figures 4–7 graphically assess some selected model performance metrics (Accuracy, Specificity, Sensitivity and AUC, respectively) for the Validation (top panels) and Prediction (bottom panels) subsets. Figures 8 and 9 show similar information for two error metrics (False Negatives and False Positives). Box-and-whiskers plots were ordered from left to right in decreasing mean AUC value (at $t_2 - t_1 = 4$ years) of the Validation runs as reference for easy comparison between figures. Similar plots for the Kappa, Precision, F1, Quantity Error and Allocation Error metrics are available in Appendix C of the Supplementary Material. Prevalence values for the simulations are listed in Table 3. Figure 10 presents a Venn diagram for visual comparison of reference changes in the prediction dataset with simulated changes by each algorithm for a 4-year time step. The results clearly demonstrate that all presented ML models overestimated the amount of change, with the Random Forest (RF) model exhibiting the least degree of overestimation compared to the others. Similar diagrams for other time steps are provided in the Appendix F.

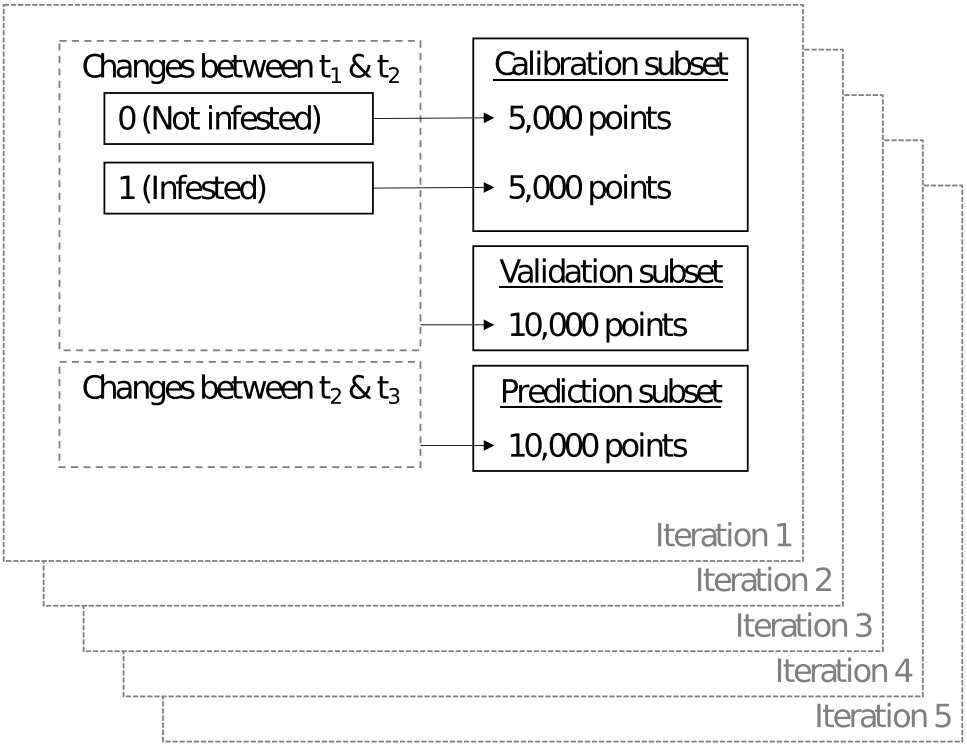

Figure 3. Selection of calibration, validation and prediction subsets from MPB data in five iterations.

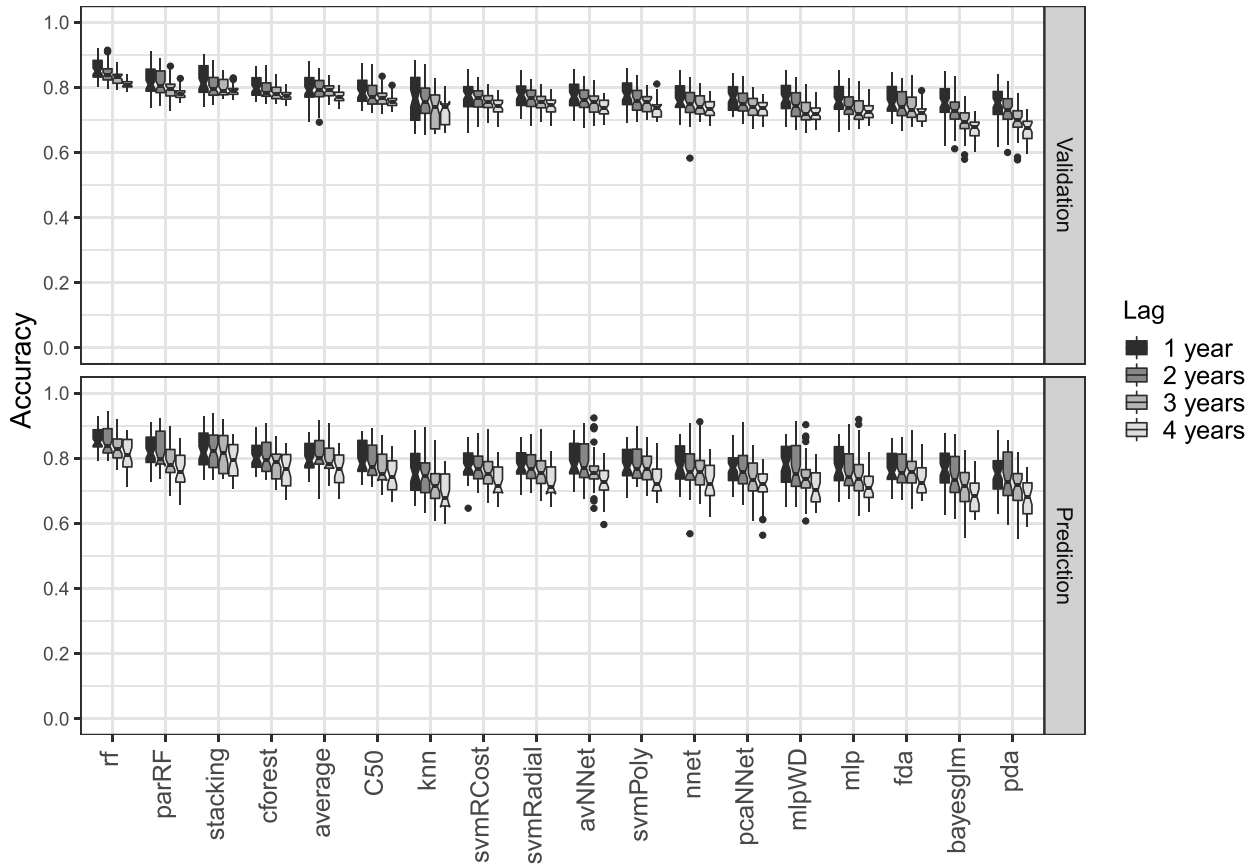

Figure 4. Box and whiskers plot of the accuracy metric for ML models. Medians are indicated by notches in the small boxes. Upper and lower limits of each box (i.e. The hinges) correspond to the first (25%) and third (75%) quartiles. Whiskers extend 1.5xIQR from the lower and upper hinges, where IQR is the inter-quartile range (i.e. 75%-25% percentiles). Outliers (i.e. points beyond both whiskers) are plotted as solid dots. For the sake of comparison, results for ML models are sorted along the horizontal axis in decreasing order of average AUC in the validation dataset at $t_2 - t_1 = 4$ years. Top panel shows results for validation points, whereas bottom panel displays results for prediction points.

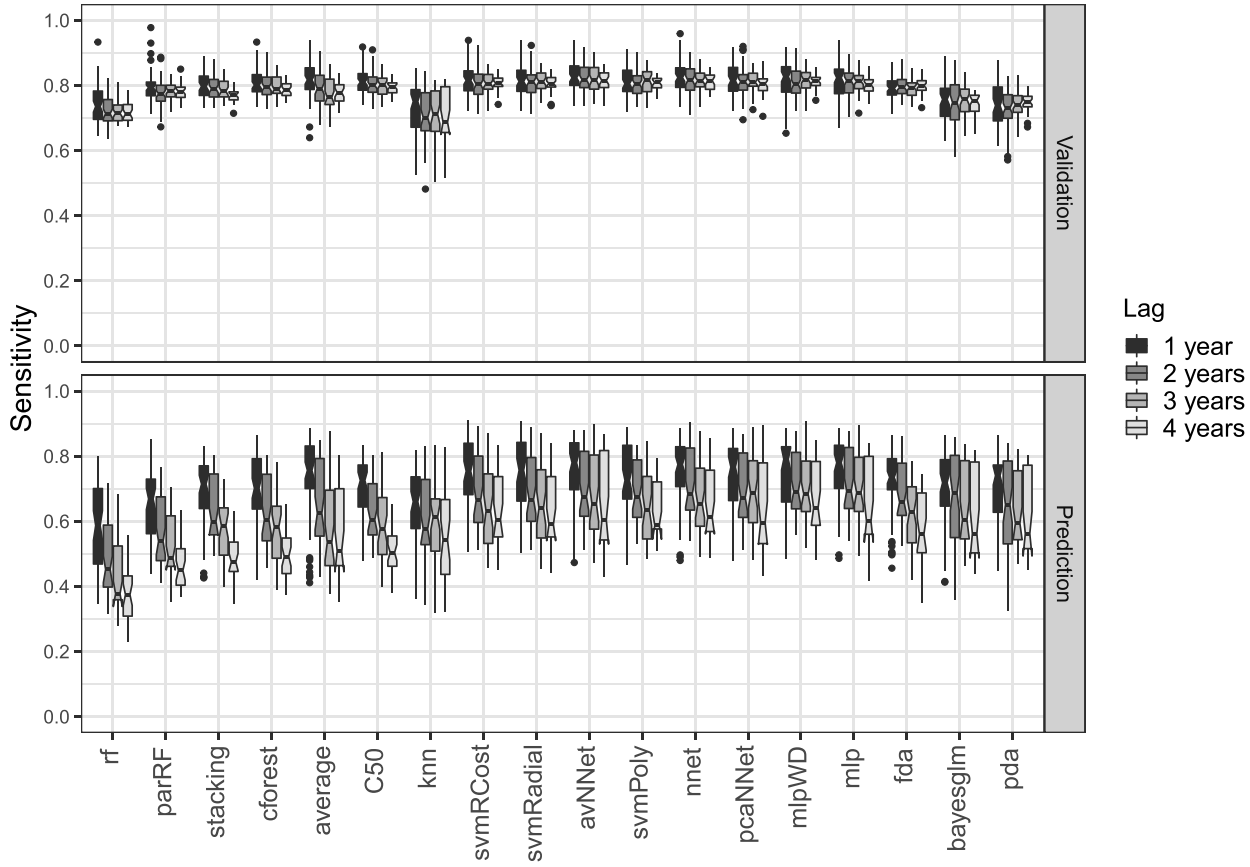

Figure 5. Same as Figure 4 for sensitivity.

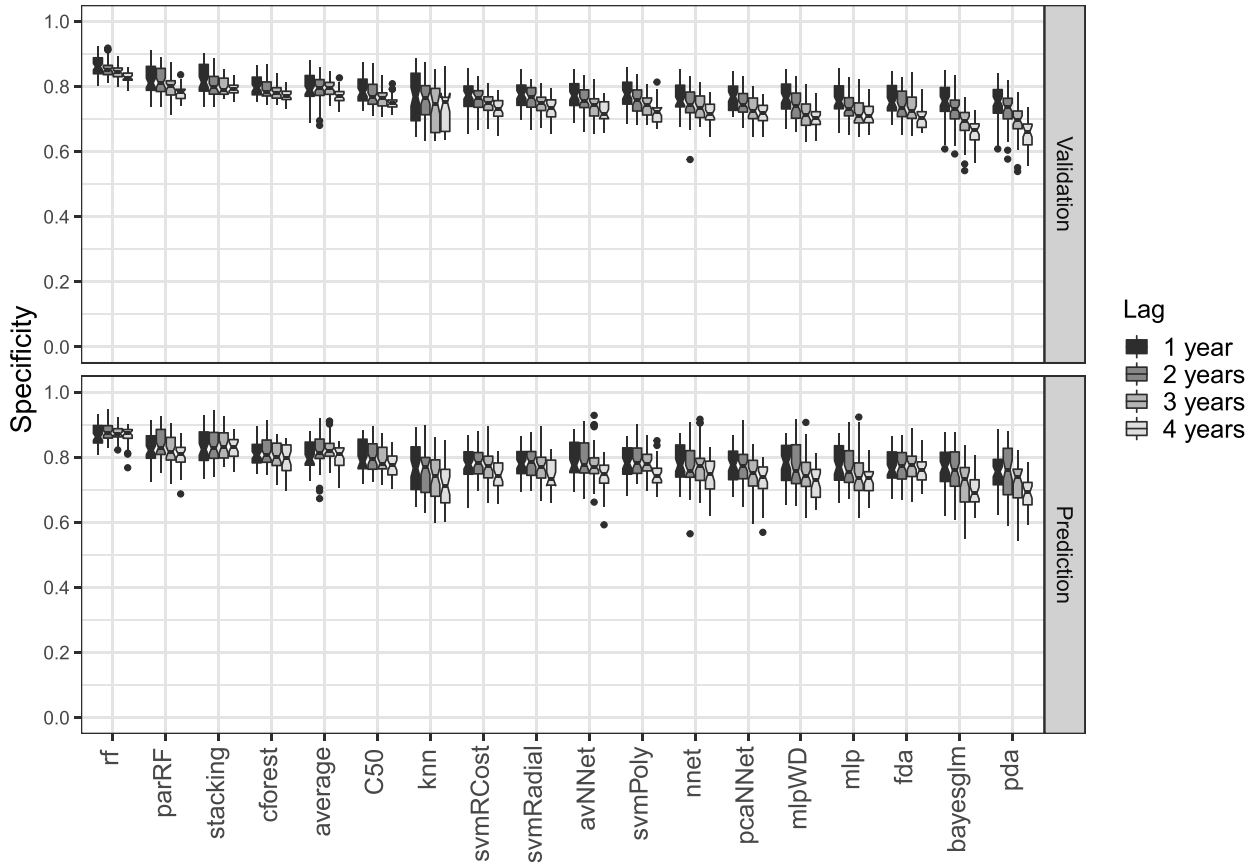

Figure 6. Same as Figure 4 for specificity.

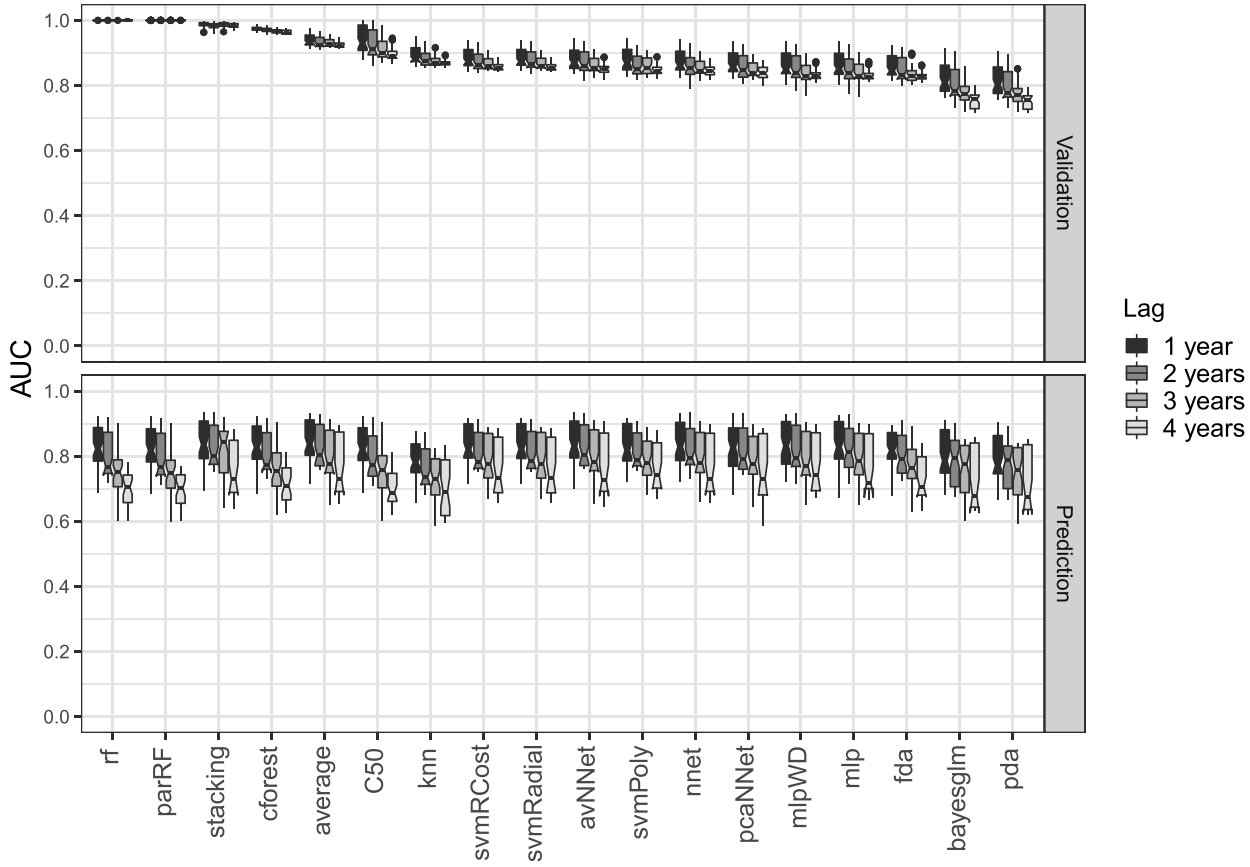

Figure 7. Same as Figure 4 for AUC.

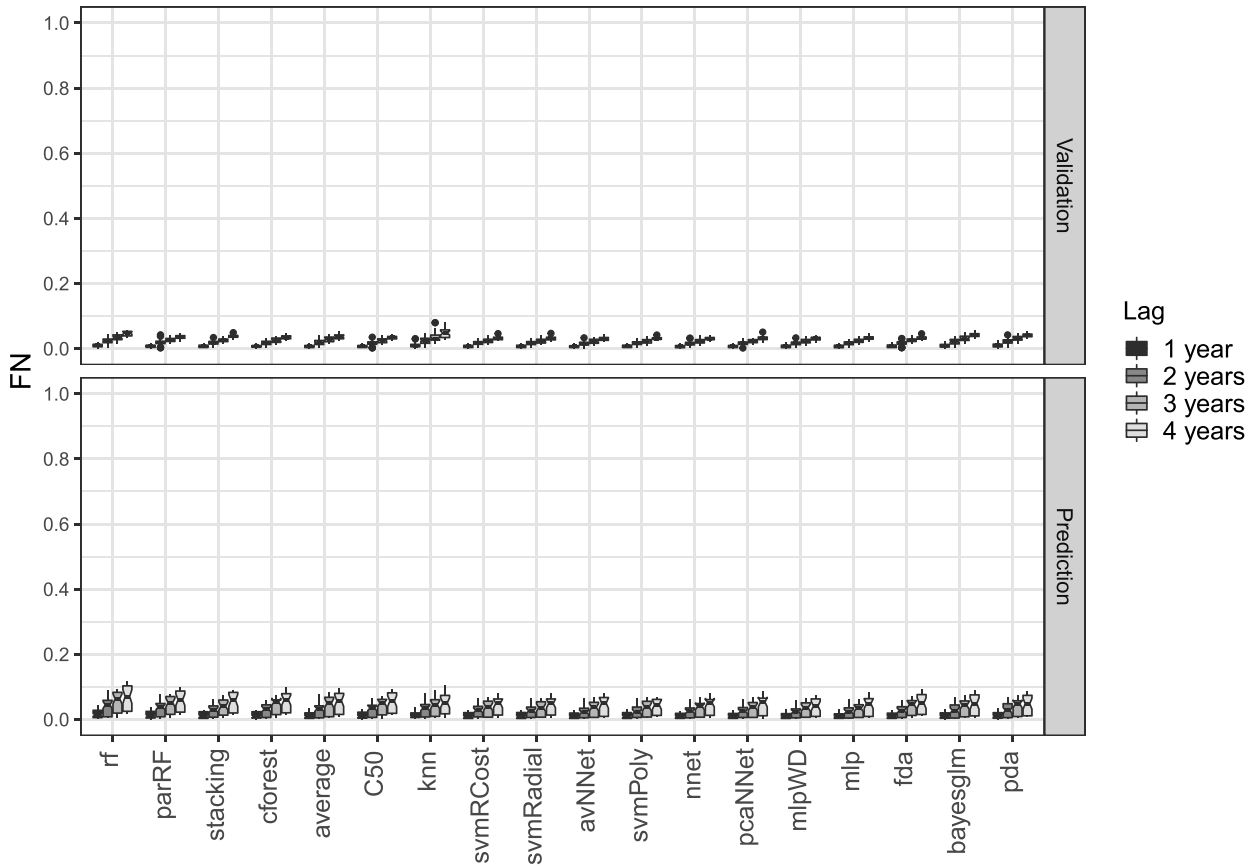

Figure 8. Same as Figure 4 for the proportion of false negative (FN) errors in subsets. Subset size was 10,000 points.

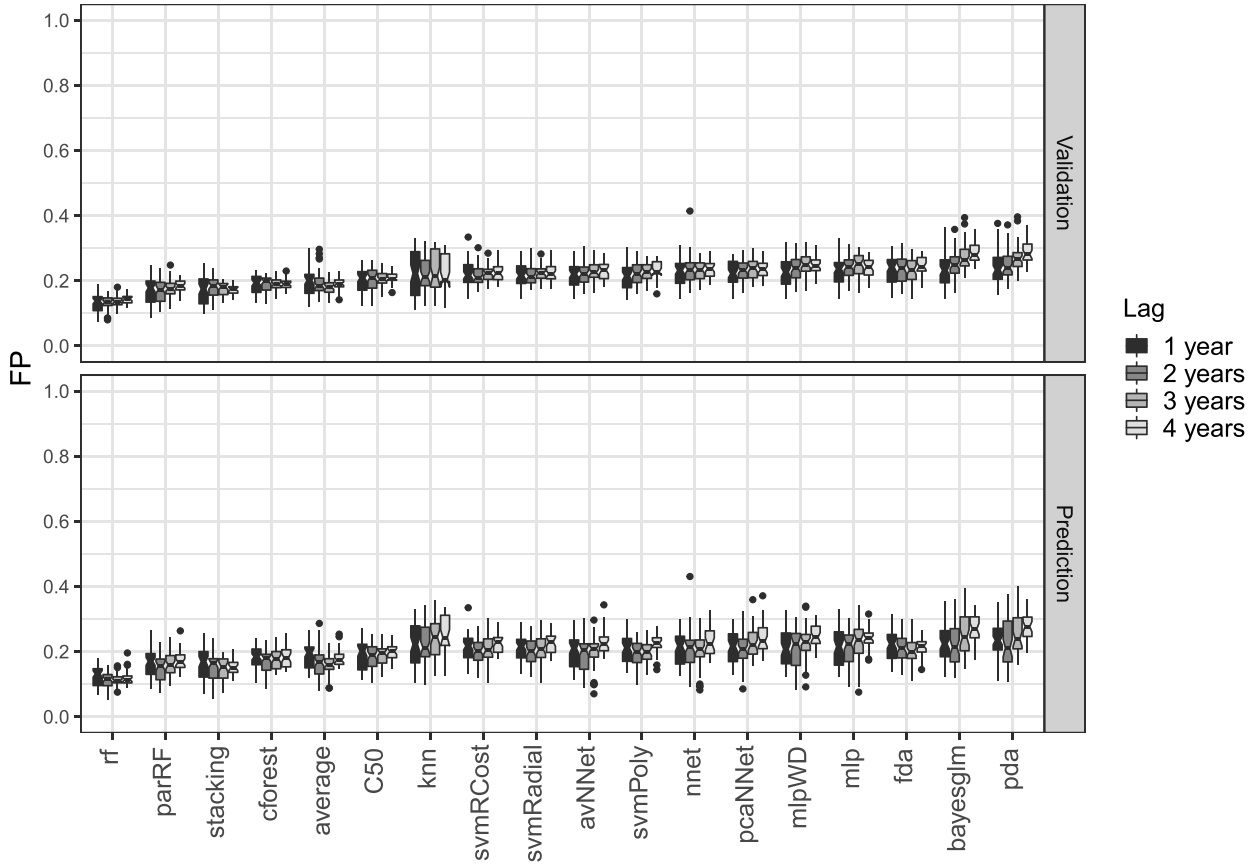

Figure 9. Same as Figure 4 for the proportion of false positive (FP) errors in subsets. Subset size was 10,000 points.

Accuracy results (Figure 4) for Validation and Prediction show the same pattern of decreasing accuracy as time lag increased. Differences between top and bottom panels in Figure 5 (Sensitivity) indicate that the ML models were able to detect correctly, on average, approximately 80% of all reference changes in the Validation dataset. On the other hand, the detection rate dropped markedly to approximately 65% for the Prediction dataset. For Accuracy and Specificity (Figures 4 and 6), however, Prediction runs showed equal or even higher values compared to Validation runs. For the AUC results (Figure 7) random forest, based models reached values very close to 1.0 for Validation, but they decreased to 0.7 for Prediction results.

False Negative errors (Figure 8) were small in validation and prediction subsets. In prediction, the median and the interquartile range of False Negative errors increased with increasing $t_2 - t_1$. False Positive errors (Figure 9) were larger than False Negatives for all models, and all time differences. Quantity Error (Figure SC4 in Supplementary Material) was smallest for random forest (rf). Allocation Error (Figure SC5 in Supplementary Material) in prediction subset increased in magnitude and range with increasing $t_2 - t_1$.

We tested the statistical significance of the observed differences between Validation and Prediction evaluation in those metrics with an Analysis of Covariance (ANCOVA) by controlling for the time step ($t_2 - t_1$) covariate. The results of the tests, illustrated in Tables D1 and D2 in Appendix D of the Supplementary Material, mostly confirmed the visual inspection of Figures 4, 5, 6 and 7, namely that AUC and Sensitivity values for the Prediction subsets were significantly lower than those for the Validation subset (a "-" sign in column VP of Tables D1 and D2). On the other hand, Specificity and to some extent Accuracy had higher values for the Prediction subset than for the validation subset. Among error metrics, False Negatives and Allocation Error increased from Validation to Prediction for most algorithms. In contrast, False Positives and Quantity Error marked a decrease from Validation to Prediction. The effect of the $t_2 - t_1$ covariate was varied: negative for Accuracy, Sensitivity, Specificity, AUC and Quantity Error, and positive for Kappa, Precision, F1, False Negatives, False Positives, and Allocation Error. In other words, the former metrics showed inferior performance and the latter metrics showed superior performance, as $t_2 - t_1$ grew from 1 to 4 years.

Table 3. Mean and standard deviation (SD) of prevalence values (%) for different $t_2 - t_1$ values.

$t_2 - t_1$ (years)	Mean prevalence	SD
1	3.5	2.0
2	7.6	3.5
3	10.6	3.9
4	13.4	4.5

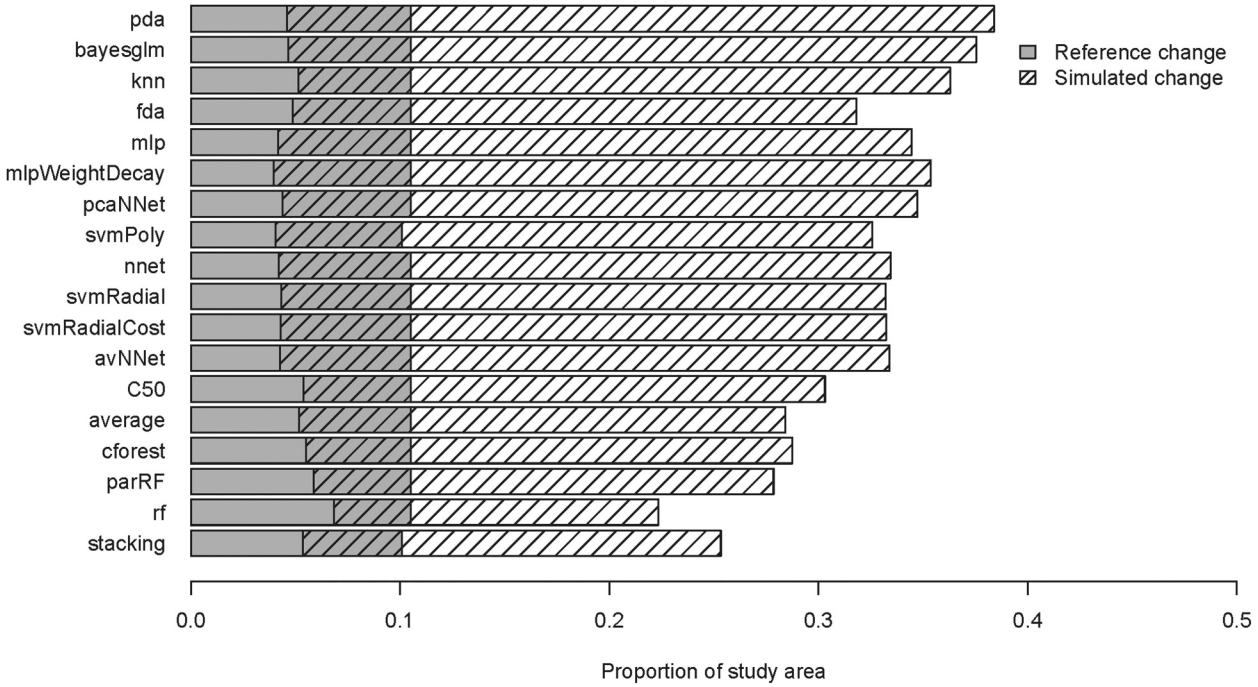

Figure 10. Venn diagrams of reference and predicted changes with 4-year lag.

3.2. Friedman ranks

Tables 4–7 show the Friedman ranks of the ML models as calculated with the Accuracy, Specificity, Sensitivity and AUC indices. Tables 8 and 9 show the calculated ranks for two error metrics (False Negatives and False Positives). Similar tables for Kappa, Precision, F1, Quantity Error and Allocation Error indices are given in Appendix E of the Supplementary Material. The procedure to calculate Friedman ranks for the simulations is described in Appendix B.

4. Discussion

4.1. Validation and prediction performance of ML models

Mathematical modeling of an ecological phenomenon, such as a spatially spreading disturbance, depends heavily on existing knowledge of the mechanisms involved in that phenomenon, as well as on available data for those mechanisms. Limited theoretical knowledge and lack of accurate data are therefore inherent challenges of modeling. Despite these problems, we tried to evaluate the performance of our models by comparing several models using a variety of metrics. We also gained insight about our models by reviewing other studies in the domain. For example, in a comparable work (Liang et al. 2014), used general linear models and a similar set of anthropogenic, biological and physical drivers to model the spread of the MPB disturbance in the Rocky Mountains. Their AUC and Accuracy results for new predicted mortality (Table 4 in that article, columns AA2) were similar to ours, though lower. Like the present study, those projections did not take into account local meteorological conditions (e.g. rain or wind).

In comparing ML models, we were particularly curious to see whether averaging or stacking of multiple ML models resulted in a substantial improvement over the best individual ML techniques. Notice that other ensemble learning techniques, as well as other ML implementations (e.g. Weka, (Frank et al. 2017; Witten et al. 2017)), may also yield different results, but those comparisons were not within the objectives of this work. It is important to have a good understanding about the performance of ensemble models because these techniques are computationally intensive and scientists may wonder whether ensemble learning classification methodologies are worth the extra calculations if improvements in performance are likely to be meager. For example, a recent study by (Hao et al. 2020) modeled presence-absence species data and concluded that ensemble techniques did not show any particular benefit over individually tuned models.

A stark reminder of how hard it is to evaluate the performance of classification models is the relative disparity between metrics. Different indices point to distinct models as the highest performers. Therefore, to interpret those performances, pertinent metrics or group of metrics should be identified. Such metrics should be consistent with the context of the study and help differentiate between high- and poor-performance models.

Table 4. Friedman ranks and average accuracy for validation and prediction datasets of all sixteen ML models, plus the two ensemble learning approaches, for $t_2 - t_1 = 4$ years.

Validation dataset			Prediction dataset		
Rank	ML classifier	Accuracy (%)	Rank	ML classifier	Accuracy (%)
1.37	rf	81.17	1.43	rf	81.30
2.45	stacking	79.04	2.45	stacking	79.39
3.71	parRF	78.12	5.07	parRF	76.77
4.57	cforest	77.50	5.11	average	76.90
5.09	average	77.29	5.80	cforest	76.26
7.00	C50	75.81	8.00	C50	74.81
9.73	svmRadial	74.23	9.26	fda	73.82
9.81	svmRadialCost	74.23	10.31	svmPoly	73.47
10.26	avNNet	73.85	10.69	avNNet	72.84
10.76	svmPoly	73.56	10.71	svmRadial	72.98
11.23	nnet	73.40	10.87	nnet	72.87
11.23	pcaNNet	73.19	10.97	svmRadialCost	72.97
11.73	knn	72.34	11.49	pcaNNet	71.37
12.29	mlp	72.76	12.43	mlp	71.87
13.26	fda	72.25	13.51	mlpWeightDecay	71.18
13.54	mlpWeightDecay	72.05	13.74	knn	69.07
17.26	bayesglm	67.18	14.69	bayesglm	68.32
17.47	pda	66.72	16.14	pda	67.51

Models are ordered from best (top) to worse (bottom) Friedman rank. Units of the accuracy column have been rescaled from [0, 1] to [0, 100] for easier reading.

Table 5. Same as Table 4 for specificity.

Validation dataset			Prediction dataset		
Rank	ML classifier	Specificity (%)	Rank	ML classifier	Specificity (%)
1.49	rf	82.99	1.43	rf	86.67
2.53	stacking	79.43	2.62	stacking	82.87
3.97	parRF	78.11	4.49	parRF	80.46
4.91	cforest	77.26	5.31	average	79.83
5.24	average	77.21	5.34	cforest	79.41
7.46	C50	75.11	7.17	C50	77.68
10.04	svmRadial	72.98	9.01	fda	76.01
10.24	svmRadialCost	72.97	10.79	svmPoly	74.81
10.31	knn	72.61	11.13	svmRadial	74.46
10.72	svmPoly	72.16	11.21	nnet	74.14
10.80	avNNet	72.43	11.29	avNNet	74.25
11.27	pcaNNet	71.80	11.36	svmRadialCost	74.41
11.49	nnet	71.86	11.57	pcaNNet	72.81
11.99	mlp	71.34	12.80	mlp	73.09
12.80	fda	70.74	13.09	knn	71.12
13.51	mlpWeightDecay	70.27	14.16	mlpWeightDecay	72.01
16.87	bayesglm	65.78	14.31	bayesglm	69.73
17.10	pda	65.13	15.70	pda	68.76

Table 6. Same as Table 4 for sensitivity.

Validation dataset			Prediction dataset		
Rank	ML classifier	Sensitivity (%)	Rank	ML classifier	Sensitivity (%)
5.49	avNNet	81.25	3.99	mlpWeightDecay	66.68
5.84	nnet	81.16	5.34	avNNet	64.87
6.09	mlpWeightDecay	81.08	5.49	mlp	64.87
6.28	svmPoly	80.68	5.60	nnet	64.32
6.56	svmRadialCost	80.65	6.70	pcaNNet	63.70
6.86	svmRadial	80.61	6.77	svmRadialCost	63.01
7.16	pcaNNet	80.38	6.99	svmRadial	62.75
7.80	mlp	80.03	7.28	svmPoly	63.08
7.89	fda	79.82	7.99	bayesglm	61.36
8.60	C50	79.41	8.20	pda	61.31
9.63	cforest	78.78	10.06	fda	57.32
10.69	parRF	78.19	11.26	knn	55.65
10.94	average	77.90	12.00	average	56.01
13.07	stacking	76.94	13.56	C50	51.25
13.91	knn	70.84	14.14	cforest	49.59
14.81	pda	74.75	15.16	stacking	49.45
15.31	bayesglm	74.26	15.91	parRF	46.44
16.89	rf	71.81	17.91	rf	37.87

Table 7. Same as Table 4 for AUC.

Validation dataset			Prediction dataset		
Rank	ML classifier	AUC (%)	Rank	ML classifier	AUC (%)
1.80	rf	100.00	4.31	avNNet	76.87
1.89	parRF	100.00	4.37	mlpWeightDecay	76.93
3.00	stacking	98.39	5.06	mlp	76.65
4.17	cforest	96.36	5.11	average	76.54
5.23	average	92.51	5.37	nnet	76.49
6.11	C50	89.65	6.69	pcaNNet	75.98
7.49	knn	87.00	7.17	svmPoly	76.21
8.83	svmRadialCost	85.81	7.46	svmRadialCost	75.69
8.91	svmRadial	85.81	7.54	svmRadial	75.68
10.40	avNNet	85.21	8.90	stacking	75.84
11.10	svmPoly	84.68	10.37	fda	73.45
11.91	nnet	84.61	12.74	bayesglm	72.95
12.86	pcaNNet	83.99	13.31	cforest	71.30
14.34	mlpWeightDecay	83.33	14.34	pda	72.09
14.49	mlp	83.25	14.54	rf	69.63
15.37	fda	82.89	14.80	C50	69.84
17.03	bayesglm	75.58	15.14	parRF	69.46
17.97	pda	75.35	16.00	knn	69.41

Table 8. Same as Table 4 for false negatives.

Validation dataset			Prediction dataset		
Rank	ML classifier	FN (%)	Rank	ML classifier	FN (%)
5.14	avNNet	3.07	3.64	mlpWeightDecay	3.95
5.50	nnet	3.07	5.00	avNNet	4.26
5.74	mlpWeightDecay	3.09	5.14	mlp	4.18
6.21	svmRadialCost	3.15	5.26	nnet	4.19
6.28	svmPoly	3.11	6.36	pcaNNet	4.39
6.51	svmRadial	3.16	6.43	svmRadialCost	4.29
6.81	pcaNNet	3.21	6.64	svmRadial	4.32
7.46	mlp	3.26	7.28	svmPoly	4.03
7.54	fda	3.28	7.64	bayesglm	4.65
8.26	C50	3.36	7.86	pda	4.60
9.29	cforest	3.47	9.71	fda	4.86
10.34	parRF	3.56	10.91	knn	5.14
10.60	average	3.62	11.66	average	5.17
13.07	stacking	3.73	13.21	C50	5.38
13.57	knn	4.74	13.80	cforest	5.50
14.47	pda	4.10	15.16	stacking	5.35
14.97	bayesglm	4.19	15.57	parRF	5.88
16.54	rf	4.60	17.57	rf	6.85

Table 9. Same as Table 4 for false positives.

Validation dataset			Prediction dataset		
Rank	ML classifier	FP (%)	Rank	ML classifier	FP (%)
1.14	rf	14.23	1.09	rf	11.85
2.53	stacking	17.24	2.62	stacking	15.25
3.63	parRF	18.32	4.14	parRF	17.35
4.57	cforest	19.04	4.97	average	17.93
4.90	average	19.08	5.00	cforest	18.24
7.11	C50	20.83	6.83	C50	19.81
9.70	svmRadial	22.61	8.67	fda	21.32
9.90	svmRadialCost	22.62	10.79	svmRadial	22.70
9.97	knn	22.92	10.79	svmPoly	22.49
10.46	avNNet	23.08	10.87	nnet	22.94
10.72	svmPoly	23.33	10.94	avNNet	22.90
10.93	pcaNNet	23.60	11.01	svmRadialCost	22.74
11.14	nnet	23.54	11.23	pcaNNet	24.24
11.64	mlp	23.98	12.46	mlp	23.94
12.46	fda	24.48	12.74	knn	25.79
13.17	mlpWeightDecay	24.87	13.81	mlpWeightDecay	24.87
16.53	bayesglm	28.64	13.97	bayesglm	27.03
16.76	pda	29.18	15.36	pda	27.88

AUC values in the validation set were above 0.8 for most models, above 0.9 for ensemble models, and even nearly 1.0 for random forest-based models. This allegedly shows that random forest-based models and ensemble models were especially successful at identifying the rules governing the training dataset. AUC values dropped to approximately 0.7 for the Prediction set, a fact which would apparently point to

a satisfactory predictive capability. Similarly, median Accuracy values were mostly above 0.7 for Validation and Prediction sets. However, this rosy picture is tempered by the results from other indices. In particular, although the Specificity metric indicates that many non-infested pixels were correctly classified as such, the Sensitivity metric shows that many infested pixels were wrongly classified as non-infested. The contradictory results for the AUC index lend support to findings in other studies (Lobo, Jiménez-Valverde, and Real 2008), which determined that AUC values can be significantly affected by low prevalence values. We further argue that such effect may also exist in studies with high prevalence values. Thus, caution should be advocated when interpreting AUC indices.

Our results also indicate that providing Accuracy alone could give a false sense of confidence in the performance of the models. One cannot deduce their true predictive ability unless performance metrics that are relevant to the objectives and context of the study are provided. Accuracy may be an inadequate metric since it weighs both true positives and true negatives in a similar way. When prediction ability is our main goal and data imbalance is a concern, Specificity and Sensitivity are far more informative than Accuracy.

It must be noted, however, that Specificity and Sensitivity, too, can be misleading if they are the only metrics used in model assessment. Consider the two hypothetical examples presented in Table 10. In both cases, Specificity (or True Negative rate) is 0.6 and Sensitivity (or True Positive rate) is 0.2. Yet this information is not enough to determine whether a model overestimates or underestimates change. In fact, comparison of False Negative and False Positive errors of the two examples clearly distinguishes between the two examples. In example 1, where prevalence is 0.50, False Negative errors are larger than False Positive errors. Thus, in example 1 change is underestimated. On the other hand, in example 2, where prevalence is 0.10, the dominant errors are False Positives, and change is overestimated. These examples show that relying on Sensitivity and Specificity, alone, is insufficient for assessing model performance. We therefore advocate that Sensitivity and Specificity, too, should be interpreted with caution, and especially with consideration of prevalence.

Next, we consider error metrics. Our results show that in model predictions, Allocation Error is small and does not substantially distinguish models from one-another (see Figure C5 and Table E5). However, values and ranges of Quantity Error were substantially larger (see Figure C4 and Table E4). This means that, firstly, Quantity Error was the major type of error that the models made, and secondly, the large range of Quantity Errors allows us to identify differences among models. Therefore, Quantity Error is a pertinent metric that we can use for comparing our ML models. A Quantity Error can be an overestimation or an underestimation of the quantity of change. In order to find out which is the case, we refer to False Positives and False Negatives (see Figures 8 and 9 and Tables 8 and 9). Our results reveal that False Negatives are relatively small in value and small in range, meaning that they do not show major differences between the models. False Positives, on the other hand, seem to be the dominant errors and show a large range that allows to distinguish models from one another. Therefore, False Positives is another pertinent metric for comparing ML models in our study.

4.2. Friedman ranks

The ML algorithms that were ranked at the top on the validation dataset performed more discreetly on the prediction datasets. Among them, the case for random-forest-based algorithms is most interesting, since random forests are often assumed to be one of the best performing ML models (Fernández-Delgado et al. 2014). Intriguingly, the range of ranks for performance metrics (but not error metrics) is systematically larger for Validation than for Prediction subsets. That is, considering performance metrics (and not error metrics), models are closer in ranks for Prediction than for Validation, suggesting that they performed more similarly in Prediction. That is not the case with error metrics. In other words, in contrast with performance metrics, our error metric ranks revealed larger differences among Prediction outputs.

Considering performance metrics alone (and not error metrics), the sums of Friedman ranks of models suggest that ensemble models (average and stacking) followed by neural network based models (avNNet and nnet) were the best machine learning models. On the lower end of the ranks were bayesglm, pda and knn. A similar conclusion is reached by considering performance and error metrics together.

The above conclusion is based on the assumption of equal importance of all metrics. A different picture emerges when metrics are studied individually. While mlp and neural-network–based models rank

Table 10. Confusion matrices for two hypothetical examples with the same sensitivity value and the same specificity value.

		Observation					Observation		
Example 1		Change	Persist	*Sum*	Example 2		Change	Persist	*Sum*
Model	Change	0.10	0.20	*0.30*	Model	Change	0.02	0.36	*0.38*
	Persist	0.40	0.30	*0.70*		Persist	0.08	0.54	*0.62*
	Sum	*0.50*	*0.50*	***1***		*Sum*	*0.10*	*0.90*	***1***

high (meaning low error rates) in AUC, Sensitivity, Allocation Error and False Negatives, rf and random forest-based models ranked high in Accuracy, Specificity, Quantity Error and False Positives. Interestingly, the ranks of stacking ensemble models are similar to random forest models in error metrics.

As discussed in the previous section, we identified Quantity Error and False Positives as the pertinent metrics for comparison of the models in our case study. As a result, we can say that rf and stacking were the high performing models, followed by parRF and average. Ensemble models performed better than most ML models except random forest models.

4.3. Implications for the evaluation of machine-learning models

The results of our study serve to highlight the inherent limitations of cross-validation techniques that rely on splitting the same dataset between calibration (aka training) and validation (aka test) subsets see e.g. (Kuhn and Johnson 2013) for an introduction to data splitting techniques). Repeated training-test cross-validation strategies may be valid and sound for problems in which the algorithms will carry out predictions under conditions that are similar to those faced during training. However, we have shown above that spatially and temporally dynamic processes like the MPB disturbance may be much harder to validate.

As we noted, model assessment is a challenging part of the modeling process. Metrics for assessment of models should be chosen and interpreted with caution and with regard to the context of the case of study and research objectives (Pontius 2022). An important characteristic of our case was that all simulations were run in a context with low prevalence. This should specially be taken into account when analyzing metrics such as Accuracy, AUC, Sensitivity and Specificity. We showed that error analysis can better describe model performance. Furthermore, direct use of components of the confusion table provided useful information that was otherwise hidden from analyses. This is in line with the recommendations of recent literature (Varga et al. 2019).

4.4. Limitations of the study

Hardware and software constraints forced us to limit the size of the datasets in our study. Although we did all our calculations with random subsets in order to make sure that the variability inherent to their relatively small size did not affect, on average, model outcome, it remains to be tested whether larger dataset may improve the ML performance. Moreover, when we drastically reduced the size of the datasets to 1,000, the results (not shown) were noisier but all trends were similar to those described above. In those calculations, however, algorithm convergence was often erratic, often forcing us to restart all over again.

We chose predictor variables based on weather rather than meteorology. The former may not be fully informative as drivers of the MPB infestation, but are relatively easy to obtain. On the other hand, variables describing local meteorological conditions are far harder to find and process, though they may indeed play a very important role in boosting or curbing the spread of the MPB disturbance (Perez and Dragicevic 2012). Therefore, future work should strive to include detailed information about, e.g. local wind conditions or cold spells into the models, which may likely enhance the performance of the models.

Finally, in random sampling, which we did due to computational limits, we lost information about adjacencies and spatial patterns. Such information, especially in a spatially spreading process like MPB infestation, could provide insight into the simulated and reference change. A particular question that can be raised in assessment of models of spatially spreading processes is about the distance of spread. Different models of the same process may simulate spread into different distances compared with one-another, and indeed in comparison with reference information. More sophisticated differences emerge where there is spatial heterogeneity in data, and where other explanatory variables than the state of each location (for example, infestation status) are considered. Therefore, assessment of models of such processes is more challenging than point-by-point comparison of simulation and reference data. New methods using naïve models of contagion for comparison with simulations of spatially spreading phenomena have been proposed in recent literature (Harati et al. 2021). Such methods add to the analysis of the confusion table and produce information on not only the count of errors, but also on where in the study area they occur, thereby enriching model assessment.

5. Conclusions

In this study, we applied sixteen machine-learning models plus two ensemble averaging procedures to MPB infestation data in British Columbia. As predictor variables driving infestation we included topographic, climatic and adjacency variables. For cross-validation, we carried out a two-fold strategy: on the one hand, we verified the results of the simulations by randomly splitting datasets between training and test subsets (so-called Validation assessment); on the other hand, we compared future projections with observations (aka Prediction assessment). We did all calculations for different MPB map sets and time differences. Seven performance metrics (six threshold-dependent and one threshold-independent) and four error metrics were used to assess

performance. To study the observed differences between Validation and Prediction metrics we ran ANCOVA tests with Validation/Prediction as fixed factor and time difference between maps $t_2 - t_1$ as covariate. In addition, Friedman rankings were computed for all simulation and performance metrics. We argue that different conclusions could be reached for different performance metrics, and therefore model assessment metrics should be selected with regard to the particular context of the case of study. We conclude that, for prediction purposes, error metrics and components of the confusion table are most helpful in understanding the ability and limitations of MPB predictive models.

Acknowledgments

The authors are thankful to Mr. Graham Hawkins from the Ministry of Forests, Lands and Natural Resource Operations of British Columbia for his support in providing data, and to the Netherlands Environmental Assessment Agency (MNP) as the owner and RIKS BV as the developer of the Map Comparison Kit for providing access to the software. The analyses were made possible thanks to provision of high-performance computing by Calcul Quebec.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, under the Discovery [grant number RGPIN/05396–2016] awarded to LP; RMH received financial support from the NEWFOREST [grant number PIRSES-GA-2013-612645] program of the European Union's Seventh Framework Programme. The Université de Montréal's International Affairs Office (IAO) also provided partial financial support through the International Partnership Development program, which allowed the collaboration between researchers from UdeM and CREAF.

Notes on contributors

Roberto Molowny-Horas earned a Ph.D. in Astrophysics from the Institute of Theoretical Astrophysics (ITA, Norway) (1994), following a BSc in Physics from Universidad de La Laguna (ULL, Spain) (1988). He held postdoctoral positions at the Instituto de Astrofísica de Canarias (IAC, Spain) (1995–1997) and the Observatoire de Paris (OBSPM, France) (1997–1999). After a four-year period in the private sector, he expanded his expertise with an MSc in Remote Sensing and GIS from the Institut d'Estudis Espacials de Catalunya (IEEC, Spain) (2003). Currently, a Research Technician at CREAF, located on the Universitat Autònoma de Barcelona (UAB, Spain) campus, he focuses on environmental sciences, applied statistics, and numerical ecology. His research has led to numerous publications in high-impact journals, addressing topics such as land cover changes in Mediterranean and Boreal landscapes, forest dynamics modeling, species interaction networks, and the impacts of insecticides on bee populations. He has been a visiting researcher at Université de Montréal (UM, Canada) and collaborates with national (CTFC, UPC) and international (UM, UNIBO, Italy) research institutions.

Saeed Harati-Asl received his Bachelor's degree in Mechanical Engineering from University of Tehran (2002) and his first Master's degree in Environmental Engineering from Azad University Science and Research Branch (2006). He gained several years of experience in development and sustainability before earning his second Master's degree in Environment and Sustainable Development from Université de Montréal (2015). He then obtained his PhD in Geography, specializing in Geomatics and Modeling, also from Université de Montréal (2022). He has completed two postdoctoral fellowships at Université de Montréal and McGill University. He is interested in developing novel applications of geomatics in sustainable energy systems.

Liliana Perez received her B.Eng in Cadastre and Geodesy from the Distrital University (Colombia) in 2000, a Master's degree in Geography from the UPTC (Colombia) in 2003, a PhD in Geography from Simon Fraser University (Canada) in 2011, followed by two postdoctoral fellowships at the University of British Columbia (Canada) in 2011 and the University of Victoria (Canada) from 2012 to 2013, where she worked on issues concerning the spatial and temporal relationships between changes in indirect indicators of biodiversity and climate change. She joined the Department of Geography at the University of Montreal (Canada) in 2013, and is currently a Full Professor and the founder and director of the Laboratory of Environmental Geosimulation (LEDGE), a multidisciplinary research group that develops innovative spatial analysis and simulation tools for various domains, such as urban, forest, and wetland ecosystems. Her research focuses on the integration of artificial intelligence (AI) and machine learning (ML) techniques into agent-based models (ABM) to create more realistic and robust models that can capture the dynamics and interactions of human and natural systems.

ORCID

Roberto Molowny-Horas http://orcid.org/0000-0003-2626-6379
Saeed Harati-Asl http://orcid.org/0000-0003-2318-9259
Liliana Perez http://orcid.org/0000-0002-6599-9893

Data availability statement

The data that support the findings of this study are available as follows: (a) Tree mortality data in raster format can be retrieved at https://www.for.gov.bc.ca/ftp/HRE/external/!publish/Web/bcmpb/Year12.; and (b) the DEM map that was used to compute some of the predictors of the models can be downloaded from https://open.canada.ca/en/open-maps.

References

Alpaydın, E. 2014. *Introduction to Machine Learning*. 3rd ed. Cambridge, Massachusetts: The MIT Press.

Axelson, J. N., R. I. Alfaro, and B. C. Hawkes. 2009. "Influence of Fire and Mountain Pine Beetle on the Dynamics of Lodgepole Pine Stands in British Columbia, Canada." *Forest Ecology and Management* 257 (9): 1874–1882. https://doi.org/10.1016/j.foreco.2009.01.047.

Axelson, J. N., R. I. Alfaro, and B. C. Hawkes. 2010. "Changes in Stand Structure in Uneven-Aged Lodgepole Pine Stands Impacted by Mountain Pine Beetle Epidemics and Fires in Central British Columbia." *The Forestry Chronicle* 86 (1): 87–99. https://doi.org/10.5558/tfc86087-1.

BC Ministry of Forests. 2015. "[Dataset] BC MPB Observed Cumulative Kill - Vol.12." https://www.for.gov.bc.ca/ftp/hre/external/!publish/web/bcmpb/Year12/.

Bleiker, K. P. 2019. "Risk Assessment of the Threat of Mountain Pine Beetle to Canada's Boreal and Eastern Pine Forests." Ottawa, ON. https://ostrnrcan-dostrncan.canada.ca/entities/publication/a50def47-511d-42e5-9265-73192c95d548 .

Bourbonnais, M. L., T. A. Nelson, and M. A. Wulder. 2014. "Geographic Analysis of the Impacts of Mountain Pine Beetle Infestation on Forest Fire Ignition." *The Canadian Geographer/Le Géographe Canadien* 58 (2): 188–202. https://doi.org/10.1111/j.1541-0064.2013.12057.x.

Bryant, C., N. R. Wheeler, F. Rubel, and R. H. French. 2017. "Kgc: Koeppen-Geiger Climatic Zones (R Software Package)." https://CRAN.R-project.org/package=kgc.

Cooke, B. J., and A. L. Carroll. 2017. "Predicting the Risk of Mountain Pine Beetle Spread to Eastern Pine Forests: Considering Uncertainty in Uncertain Times." *Forest Ecology and Management* 396 (July): 11–25. https://doi.org/10.1016/j.foreco.2017.04.008.

Corbett, L. J., P. Withey, V. A. Lantz, and T. O. Ochuodho. 2016. "The Economic Impact of the Mountain Pine Beetle Infestation in British Columbia: Provincial Estimates from a CGE Analysis." *Forestry* 89 (1): 100–105. https://doi.org/10.1093/forestry/cpv042.

Deane-Mayer, Z. A., and J. E. Knowles. 2023. "Caretensemble: Ensembles of Caret Models (R Software Package)." https://zachmayer.github.io/caretEnsemble/.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *The Journal of Machine Learning Research* 15 (90): 3133–3181. https://jmlr.org/papers/v15/delgado14a.html.

Frank, E., M. A. Hall, and I. H. Witten. 2017. "The WEKA Workbench." In *Data Mining: Practical Machine Learning Tools and Techniques*." In editor by I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, 553–571. 4th ed. Cambridge, MA: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-804291-5.00024-6.

Government of British Columbia. 2023. "vri - HISTORICAL Vegetation Resource Inventory (2002–2021)." https://open.canada.ca/data/en/dataset/02dba161-fdb7-48ae-a4bb-bd6ef017c36d .

Government of Canada. 2024. "Open Government Portal." https://open.canada.ca/en/open-maps.

Hao, T., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2020. "Testing Whether Ensemble Modelling is Advantageous for Maximising Predictive Performance of Species Distribution Models." *Holarctic Ecology* 43 (4): 549–558. https://doi.org/10.1111/ecog.04890.

Harati, S., L. Perez, and R. Molowny-Horas. 2020. "Integrating Neighborhood Effect and Supervised Machine Learning Techniques to Model and Simulate Forest Insect Outbreaks in British Columbia, Canada." *Forests* 11 (11): 1–23. https://doi.org/10.3390/f11111215.

Harati, S., L. Perez, R. Molowny-Horas, and R. Gilmore Pontius. 2021. "Validating Models of One-Way Land Change: An Example Case of Forest Insect Disturbance." *Landscape Ecology* 36 (10): 2919–2935. https://doi.org/10.1007/s10980-021-01272-0.

Haughian, S. R., P. J. Burton, S. W. Taylor, and C. Curry. 2012. "Expected Effects of Climate Change on Forest Disturbance Regimes in British Columbia." *BC Journal of Ecosystems and Management* 13 (1): 1–24. https://doi.org/10.22230/jem.2012v13n1a152.

He, H., and E. A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

Ho, S. Y., S. Tan, C. Chau Sze, L. Wong, and W. Wen Bin Goh. 2021. "What Can Venn Diagrams Teach Us About Doing Data Science Better?" *International Journal of Data Science and Analytics* 11 (1): 1–10. https://doi.org/10.1007/s41060-020-00230-4.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. "An Introduction to Statistical Learning." In *Springer Texts in Statistics*, Vol. 103. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Kastridis, A., D. Stathis, M. Sapountzis, and G. Theodosiou. 2022. "Insect Outbreak and Long-Term Post-Fire Effects on Soil Erosion in Mediterranean Suburban Forest." *The Land* 11 (6): 911. https://doi.org/10.3390/land11060911.

Klenner, W., R. Walton, A. Arsenault, and L. Kremsater. 2008. "Dry Forests in the Southern Interior of British Columbia: Historic Disturbances and Implications for Restoration and Management." *Forest Ecology and Management* 256 (10): 1711–1722. https://doi.org/10.1016/j.foreco.2008.02.047.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, et al. 2023. "Caret: Classification and Regression Training (R Software Package)." https://github.com/topepo/caret/ .

Lemmen, D. S., F. J. Warren, J. Lacroix, and E. Bush, eds. 2008. *From Impacts to Adaptation: Canada in a Changing Climate 2007*, 448. Ottawa, ON: Natural Resources Canada (NRC).

Liang, L., T. J. Hawbaker, Y. Chen, Z. Zhu, and P. Gong. 2014. "Characterizing Recent and Projecting Future Potential Patterns of Mountain Pine Beetle Outbreaks in the Southern Rocky Mountains." *Applied Geography* 55:165–175. https://doi.org/10.1016/j.apgeog.2014.09.012.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." *Global Ecology and Biogeography* 17 (2): 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

MacLean, D. A. 2016. "Impacts of Insect Outbreaks on Tree Mortality, Productivity, and Stand Development." *The Canadian Entomologist* 148 (S1): S138–59. https://doi.org/10.4039/tce.2015.24.

Mahdizadeh Gharakhanlou, N., and L. Perez. 2024. "From Data to Harvest: Leveraging Ensemble Machine Learning for Enhanced Crop Yield Predictions Across Canada Amidst Climate Change." *Science of the Total Environment* 951:175764. November. https://doi.org/10.1016/j.scitotenv.2024.175764.

McCullough, D. G., R. A. Werner, and D. Neumann. 1998. "Fire and Insects in Northern and Boreal Forest Ecosystems of North America." *Annual Review of Entomology* 43 (1): 107–127. https://doi.org/10.1146/annurev.ento.43.1.107.

McGillivray, B. 2011. *Geography of British Columbia: People and Landscapes in Transition*. 3rd ed. Vancouver, BC: UBC Press.

O'Brien, R. M., and F. Zhou. 2018. "A Consistent and General Modified Venn Diagram Approach That Provides Insights into Regression Analysis." *Public Library of Science ONE* 13 (5): e0196740. https://doi.org/10.1371/journal.pone.0196740.

Patel, J., J. Katan, L. Perez, and R. Sengupta. 2021. "Transferring Decision Boundaries Onto a Geographic Space: Agent Rules Extracted from Movement Data Using Classification Trees." *Transactions in GIS* 25 (3): 1176–1192. https://doi.org/10.1111/tgis.12770.

Patriquin, M. N., A. M. Wellstead, and W. A. White. 2007. "Beetles, Trees, and People: Regional Economic Impact Sensitivity and Policy Considerations Related to the Mountain Pine Beetle Infestation in British Columbia, Canada." *Forest Policy and Economics* 9 (8): 938–946. https://doi.org/10.1016/J.FORPOL.2006.08.002.

Pelz, K. A., and F. W. Smith. 2012. "Thirty Year Change in Lodgepole and lodgepole/Mixed Conifer Forest Structure Following 1980s Mountain Pine Beetle Outbreak in Western Colorado, USA." *Forest Ecology and Management* 280 (September): 93–102. https://doi.org/10.1016/J.FORECO.2012.05.032.

Perez, L., and S. Dragicevic. 2012. "Landscape-Level Simulation of Forest Insect Disturbance: Coupling Swarm Intelligent Agents with Gis-Based Cellular Automata Model." *Ecological Modelling* 231 (April): 53–64. https://doi.org/10.1016/j.ecolmodel.2012.01.020.

Petersen, B., and D. Stuart. 2014. "Explanations of a Changing Landscape: A Critical Examination of the British Columbia Bark Beetle Epidemic." *Environment and Planning A* 46 (3): 598–613. https://doi.org/10.1068/a4672.

Pontius, R. G., Jr. 2022. "Metrics That Make a Difference." In *Advances in Geographic Information Science*, Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-70765-1.

Raffa, K. F., B. H. Aukema, B. J. Bentz, A. L. Carroll, J. A. Hicke, M. G. Turner, and W. H. Romme. 2008. "Cross-Scale Drivers of Natural Disturbances Prone to Anthropogenic Amplification: The Dynamics of Bark Beetle Eruptions." *BioScience* 58 (6): 501–517. https://doi.org/10.1641/B580607.

Robbins, J. 2008. "Bark Beetles Kill Millions of Acres of Trees in West." *New York Times, November 18, 2008*.

Safranyik, L., and A. L. Carroll. 2007. "The Biology and Epidemiology of the Mountain Pine Beetle in Lodgepole Pine Forests." In *The Mountain Pine Beetle: A Synthesis of Biology, Management, and Impacts on Lodgepole Pine*, edited by L. Safranyik and W. R. Wilson, 3–66. Victoria, British Columbia: Canadian Forest Service.

Strohm, S., M. L. Reid, and R. C. Tyson. 2016. "Impacts of Management on Mountain Pine Beetle Spread and Damage: A Process-Rich Model." *Ecological Modelling* 337 (October): 241–252. https://doi.org/10.1016/J.ECOLMODEL.2016.07.010.

Varga, O. G., R. G. Pontius Jr., S. K. Singh, and S. Szabó. 2019. "Intensity Analysis and the Figure of Merit's Components for Assessment of a Cellular Automata – Markov Simulation Model." *Ecological Indicators* 101 (January): 933–942. https://doi.org/10.1016/j.ecolind.2019.01.057.

Witten, I. H., E. Frank, M. A. Hall, C. J. Pal, and M. Data. 2017. "Practical machine learning tools and techniques." In *Data mining*, 403–413. Vol. 2. Amsterdam, The Netherlands: Elsevier Publishers.

Zhao, L. Q., S. Dragićević, S. Balram, and L. Perez. 2025. "Assessing the Number of Criteria in gis-Based Multicriteria Evaluation: A Machine Learning Approach." *Geographical Analysis*. https://doi.org/10.1111/gean.70004.