



Proceeding Paper

Prediction and Classification of Flood Susceptibility Based on Historic Record in a Large, Diverse, and Data Sparse Country †

Heather McGrath *  and Piper Nora Gohl

Natural Resources Canada, Ottawa, ON K1A 0E4, Canada

* Correspondence: heather.mcgrath@nrcan-rncan.gc.ca

† Presented at the 7th International Electronic Conference on Water Sciences, 15–30 March 2023; Available online: <https://ecws-7.sciforum.net/>.

Abstract: The emergence of Machine learning (ML) algorithms has shown competency in a variety of fields and are growing in popularity in their application to geospatial science issues. Most recently, and notably, ML algorithms have been applied to flood susceptibility (FS) mapping. Leveraging high-power computing systems and existing ML algorithms with national datasets of Canada, this project has explored methods to create a national FS layer across a geographically large and diverse country with limited training data. First, approaches were considered on how to generate a map of FS for Canada at two different levels, (i) national, which combined all training data into one model, and (ii) regional, where multiple models were created, based on regional similarities, and the results were mosaicked to generate a FS map. The second experiment explored the predictive capability of several ML algorithms across the geographically large and diverse landscape. Results indicate that the national approach provides a better prediction of FS, with 95.7% of the test points, 91.5% of the pixels in the training sites, and 89.6% of the pixels across the country correctly predicted as flooded, compared to 65.5%, 80.6% and 75.6%, respectively, in the regional approach. ML models applied across the country found that support vector machine (svmRadial) and Neural Network (nnet) performed poorly in areas away from the training sites, while random forest (parRF) and Multivariate Adaptive Regression Spline (earth) performed better. A national ensemble model was ultimately selected as this blend of models compensated for the biases found in the individual models in geographic areas far removed from training sites.

Keywords: flood susceptibility; Canada; machine learning; flood priority setting



Citation: McGrath, H.; Gohl, P.N. Prediction and Classification of Flood Susceptibility Based on Historic Record in a Large, Diverse, and Data Sparse Country. *Environ. Sci. Proc.* **2023**, *25*, 18. <https://doi.org/10.3390/ECWS-7-14235>

Academic Editor: Athanasios Loukas

Published: 16 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Flooding may occur any time of the year in Canada; however, the risk is generally higher in the spring due to heavy rainfall coupled with a rapidly melting snowpack or ice jamming [1]. The development of flood hazard maps is a provincial responsibility. Thus, to date, the areas selected for mapping are determined provincially, based on their selection criteria [2]. Hazard maps generally cover a geographically small region using high-resolution datasets, which provides high-quality maps for these targeted areas. The limitation to this approach is that only those selected areas have flood maps generated, leaving many communities and large geographic areas without flood maps or indicators of their flood risk.

The emergence of Machine learning (ML) algorithms has shown competency in a variety of fields and are growing in popularity in their application to geospatial science issues. Most recently, and notably, they have been applied to FS mapping [3–6]. There are hundreds of ML algorithms available, which can be generally categorized into groups of classification or regression and may belong to different families, such as neural networks, boosting, random forests, generalized linear models, etc. [7]. These models come from different backgrounds, e.g., statistics (generalized linear models), artificial intelligence

and data mining (decision trees), and clustering (K-means), while others represent a connectionist approach (neural networks) [7]. Tools, such as Classification And Regression (caret) library in R, provide an easy mechanism to execute and compare results from a large number of classifiers [8]. The pre-eminence of Random Forest (RF) in solving problems, such as FS, are well supported [3,7,9–12].

In the literature, there is not a consensus on whether a single model or an ensemble model is more appropriate, and different disciplines report varying outcomes. Ensemble learning and hybrid ML methods have been shown to significantly improve performance, especially in cases of high dimensional data [13–15]. Both ensembles and hybrid approaches employ a type of information fusion through differing approaches [14].

Most of the existing literature which has explored ML as applied to the FS problem have covered relatively small geographic regions, and often, single watersheds. Few have explored how these models perform over vast geographic and meteorologically diverse environments. In this research, two questions are explored to create a national FS map for Canada: (i) can a single ML model outperform a collection of regional models across a geographically large and diverse region, and (ii) how model performance is affected with increasing distance from the training sites.

2. Materials and Methods

Following common practices in ML workflows, datasets that may contribute to flooding were first identified, and historic flood events were used as labeled data and split into two classes: flooded/not flooded. The data was pre-processed to scale and centre or set as factors for the nominal datasets. It was split into training and test sets (70/30) using the Classification And Regression Training (caret) create data partition function to create a balanced split [11]. Next, Variable Selection using Random Forest (VSURF) was run to identify the relevant (important) datasets (factors) by finding the model and associated datasets with the smallest out-of-bag (OOB) error. This sub-set of factors was then refined by testing for multicollinearity among the independent variables by assessing the Variance Inflation Factor (VIF) and Pearson correlation coefficient to come up with a final list of important factors. Then, the selected factors were run through a variety of ML models to determine performance. Selection of models include those commonly cited in the literature, found to perform well in FS problems, and aligned with findings from [8] in their comprehensive study. K-fold cross-validation of the training points ($K = 3$) was repeated for five resampling iterations and the average score was recorded. All ML models were accessed from the caret library in R. The train function in caret was used, which sets up a grid of tuning parameters for a large number of classification and regression algorithms [7]. These control parameters prescribe the computational nuances of the train function. The test data was then run through the trained model and the classification and FS prediction were saved. These results were then analyzed in a confusion matrix, summary statistics were calculated, and the results compared to a historic event database, Figure 1.

2.1. Single Model vs. Multi-Model

As Canada spans a geographically large area and contains a diverse set of climate characteristics, a regional approach was first considered. This approach would see several models developed. National datasets in Canada have been developed based on a variety of frameworks, such as regional, physiological, and ecozone approaches. Commonly accepted in Canada is adopting the National Ecological Framework of Canada, which provides standard ecological units [16]. This classification system “incorporates all major components of ecosystems: air, water, land, and biota” [16]. A second framework considers physiographic regions and is comprised of the shield and borderlands, each sub-divided into additional groupings [17]. As undertaken by [3], in their US work of flood risk, drainage areas were considered as natural region boundaries. In Canada, there are 11 major drainage areas.

In the national approach, all training data were processed together to generate a single model, which was then applied across the whole of the Country.

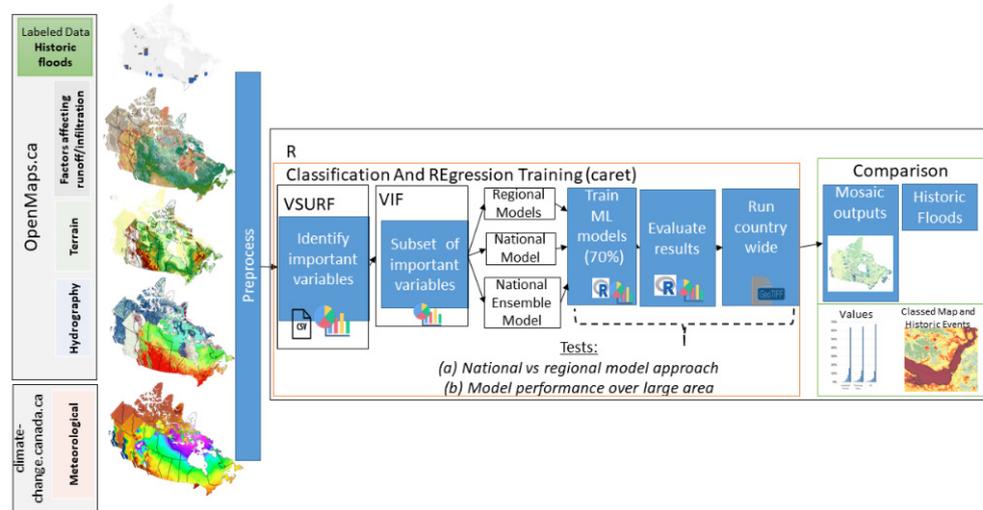


Figure 1. Methodology and tests run to create a flood susceptibility dataset in Canada.

2.2. Single Model vs. Ensemble over Large Distance

Most of the existing literature applying ML to the FS problem have tested and applied the results over a relatively small geographic areas, which are not dissimilar from the data found in the training/test set. Given the vast region of interest and the variety in land use, geography, climate, etc., several single ML models, which encompass a variety of classification and regression algorithms, were tested to evaluate their performance, Table 1. In addition, an ensemble model was developed and evaluated. Qualitative and quantitative analysis were performed to evaluate how the different ML models perform over Canada as a whole.

Table 1. Machine learning algorithms tested, C = classification, R = Regression).

Category	Algorithm	Acronym	Type
Decision Trees (DT)	C5.0	C5.0	C
Random Forest (RF)	Random Forest	RF	C, R
	Parallel Random Forest	parRF	C, R
	Regularized Random Forest	RRF	C, R
	Random Forest	ranger	C, R
Boosting (B)	eXtreme Gradient Boosting	xgbDART	C, R
	eXtreme Gradient Boosting	xgbTree	C, R
Support Vector Machines (SVM)	Support Vector Machines with Radial Basis Function Kernel	svmRadial	C, R
	Support Vector Machines with Polynomial Kernel	svmPoly	C, R
	Support Vector Machines with Radial Basis Function Kernel	svmRadialCost	C, R
Neural Networks (NN)	Model Averaged Neural Network	avNNNet	C, R
	Neural Networks with Feature Extraction	pcaNNNet	C, R
	Neural Network	nnet	C, R
	Multi-Layer Perceptron	mlp	C, R
Multivariate Adaptive Regression Splines	Multivariate Adaptive Regression Spline (MARS)	earth	C, R
Ensemble	Ensemble of Random Forest, Support Vector Machine, Boosting, and Neural Networks	parRF, nnet, svmPoly, xgbTree, pcaNNNet	C, R

2.3. ML Models Tested

The single model approach focused on Random Forest, due to its superior performance as found in the literature, [3,7,9–12]. Additional models were considered for the ensemble method and synthesized via a generalized linear model (GLM).

2.4. Validation

Analysis of the results are per confusion matrix, including several measures: Accuracy, Kappa, Sensitivity, Specificity, and F1 score.

$$\text{Accuracy } Acc = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$\text{Sensitivity } Sn = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity } Sp = \frac{TN}{TN + FP} \quad (3)$$

$$\text{F1 score } F = 2 \times \frac{P \times Sn}{P + Sn} \quad (4)$$

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (5)$$

where TP is a true positive prediction, TN is true negative, FP is a false positive, and TN is true negative prediction. The receiver operating characteristics (ROC) area under the curve (AUC), which offers an overall assessment of the model at all classification levels, was also evaluated.

2.5. Data and Study Area

Data used in this project are from nationally available datasets available from a variety of federal government organizations. All data used in this project is publicly accessible and can be found at open.canada.ca (accessed on 10 September 2021) or climate.weather.gc.ca/ (accessed on 20 August 2021). Five sites were selected across Canada, all with a history of flooding and which experienced flooding in the past ten years. The sites include southern British Columbia (BC), which experienced flooding due to atmospheric river in 2021, and flooding south of Lake Athabasca in northern Alberta (AB) due to heavy rainfall. Southern Manitoba (MB), Ontario (ON), and New Brunswick (NB) all have flood events due to the spring freshet along the Red River, Ottawa River, and Saint John River, respectively.

3. Results

In this section, results are presented for (i) single and multi-region model approaches, (ii) single model results and ensemble model across Canada, and (iii) comparison of FS prediction to historic flood events database.

3.1. Single and Multi-Region Model

Results from parallel Random Forest model (parRF) are shown in Table 2 for the multi-region and national approaches using local, regionally important variables and a compiled national set of important variables. For most of the measures in each of the study areas, the results are nearly identical between the local, self-selected factors and the national list of factors. The ON region is the exception, where an increase in model performance was found in all measures using the national list of important factors. Notably, in ON, there was an increase in accuracy from 0.89 to 0.92 in overall accuracy and increase by 0.07 in kappa and specificity when the national list of factors was used. In BC, there is an increase of 0.01 in specificity and decrease of 0.02 in both sensitivity and F1 score between the national and

local list of factors. The average of the regional models found a slight increase in accuracy, kappa, and specificity when the national factor list was used.

Table 2. Metrics of the individual models and national model, parallel random forest results shown.

	Regional Models (parRF) Local Variables/National Variables						Single Model
	BC	AB	MB	ON	NB	Average	National Model
Accuracy	0.96/0.96	0.94/0.94	0.82/0.82	0.89/0.92	0.99/0.99	0.91/0.93	0.92
Kappa	0.93/0.93	0.88/0.88	0.64/0.64	0.77/0.84	0.97/0.98	0.84/0.85	0.83
Sensitivity	0.95/0.93	0.91/0.91	0.79/0.79	0.91/0.92	0.98/0.98	0.91/0.91	0.91
Specificity	0.98/0.99	0.97/0.97	0.85/0.85	0.85/0.92	0.99/1.0	0.93/0.95	0.9
Precision	0.98/0.98	0.98/0.97	0.81/0.81	0.89/0.92	0.99/1.0	0.91/0.90	0.9
F1	0.95/0.93	0.91/0.91	0.79/0.79	0.91/0.92	0.98/0.98	0.91/0.91	0.91
AUC-ROC	0.97/0.97	0.96/0.96	0.86/0.86	0.92/0.93	0.99/0.99	0.94/0.94	0.97

3.2. Single Model vs. Ensemble over Large Distance

The National model approach was run on all the models listed in Table 1. The four RF models all perform well, with accuracy of 0.91 or 0.92, and ROC-AUC of 0.96–0.97, respectively. Similar results are found with the two boosting models, xgbDART and xgbTree, and the C5.0 decision tree model. The poorest results come from the mlp model, with an accuracy of 0.76, kappa of 0.52, and ROC-AUC 0.85. The MARS earth model has an accuracy of 0.82 and kappa of 0.64, putting it at the lower end of the performance scale. The results of the models produce results in the range of 0 to 1, where 0 is no flooding and 1 is flooding/wet pixels, and are multiplied by 100 to avoid storing float values in the final dataset.

To compare how the different ML models performed across the country, in areas distinct from the training sites, the national approach was applied to several ML models to evaluate the resultant map, Figure 2. As a reference, the extent of historic flooding for which there is a digital record, is shown in Figure 2a. The RF and earth models present, at a national scale, are somewhat similar susceptibility maps, though the earth model has higher predictions, especially in Nunavut (NU), Northwest Territories (NT), into Saskatchewan (SK) and Manitoba (MB). The earth model computes very high susceptibility values along the western shore of Hudson Bay. This is not found in the other models, nor is it present in the historic record. The RF model, in northern NU, and NT computes a relatively stable prediction without much variation between the islands nor along the shorelines in NU. There are a few spots in western Yukon (YT), which have higher predictions, and this corresponds to the location of meteorological stations. All models capture higher susceptibility to flooding along the southern borders of SK and MB, which aligns with the historic record. The nnet model appears to do well with predictions where the training data exists and provides relatively low susceptibility values in training-sparse regions: northern Quebec (QC), NU, and around the Canadian Rockies have very low susceptibility values. The svmRadial model shows peculiar ‘rings’ with higher predictions at the outer edges, covering most of NU and eastern NT. These areas may be more diverse from the labelled data than any others, and it is clear the svmRadial model had challenges.

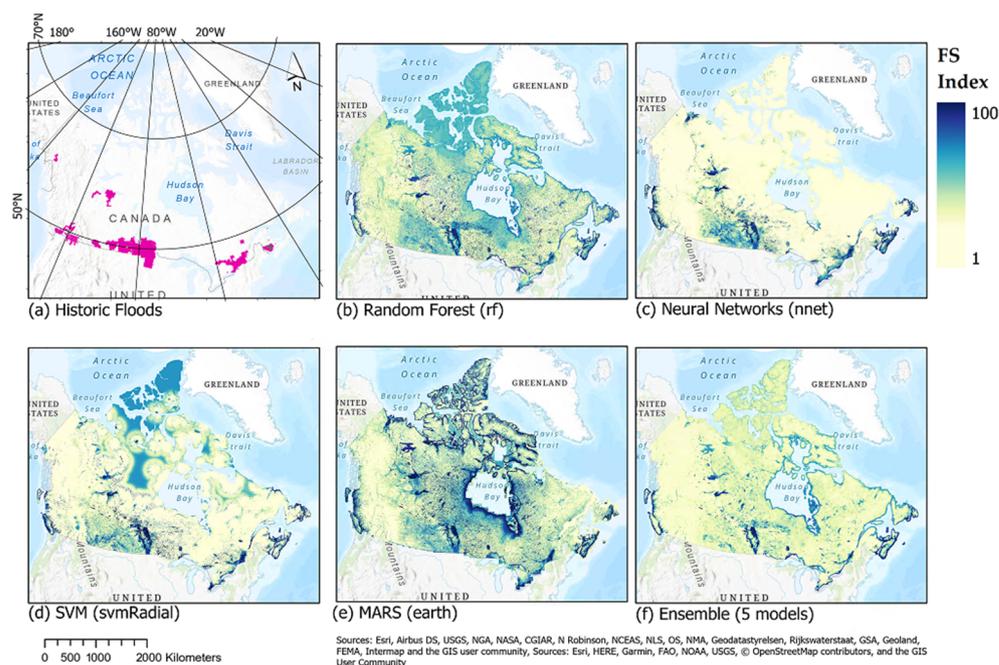


Figure 2. (a) Historic flooding in Canada (from EGS Flood Archive and national flood hazard data layer), and results of different single ML models using national approach, (b–f) ensemble result.

4. Conclusions

Developing a FS map across a nation as geographically large and diverse as Canada, presents several challenges. In this work, ML algorithms and publicly available national datasets were included to map FS and identify regions more prone to flooding. Testing if a single national model outperformed a regional multi-region model mosaic found that the single national model produced better predictions. However, when a single ML model was extrapolated across the whole of Canada, there were limitations found in several models, including SVM, NN, MLP, and RF. An ensemble approach ultimately produced the best FS map, in comparison to historic flood maps, even though the statistics from the confusion matrix found the ensemble was not the best performer with accuracy and ROC-AUC of 0.89 and kappa 0.78. The resultant dataset provides the first continuous, national picture of flood susceptibility in Canada, with the intended use to support identification and priority setting of flood hazard mapping project and for flood awareness communication.

Author Contributions: H.M.: conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; supervision; validation; visualization; roles/writing—original draft; writing—review and editing. P.N.G.: data curation; writing—review. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available: <https://open.canada.ca/data/en/dataset/df106e11-4cee-425d-bd38-7e51ac674128> (accessed on 15 June 2022).

Acknowledgments: Natural Resources Canada Contribution Number: 20220100.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Canada, P.S. Floods. Available online: <https://www.publicsafety.gc.ca/cnt/mrgnc-mngmnt/ntrl-hzrds/fl-d-en.aspx> (accessed on 7 November 2022).
2. Government of Canada, N.R.C. GEOSCAN Search Results: Fastlink. Available online: <https://geoscan.nrcan.gc.ca/starweb/geoscan/servlet.starweb?path=geoscan/fulle.web&search1=R=308128> (accessed on 4 May 2022).
3. Collins, E.L.; Sanchez, G.M.; Terando, A.; Stillwell, C.C.; Mitasova, H.; Sebastian, A.; Meentemeyer, R.K. Predicting Flood Damage Probability across the Conterminous United States. *Environ. Res. Lett.* **2022**, *17*, 034006. [CrossRef]
4. Dodangeh, E.; Choubin, B.; Eigdir, A.N.; Panahi, M.; Shamshirband, S.; Mosavi, A. Integrated Machine Learning Methods with Resampling Algorithms for Flood Susceptibility Prediction-ScienceDirect. Available online: <https://www.sciencedirect.com/science/article/pii/S0048969719359789> (accessed on 13 May 2022).
5. Shafizadeh-Moghadam, H.; Valavi, R.; Shahabi, H.; Chapi, K.; Shirzadi, A. Novel Forecasting Approaches Using Combination of Machine Learning and Statistical Models for Flood Susceptibility Mapping. *J. Environ. Manag.* **2018**, *217*, 1–11. [CrossRef] [PubMed]
6. Zhao, G.; Pang, B.; Xu, Z.; Tu, T. Mapping Flood Susceptibility in Mountainous Areas on a National Scale in China-ScienceDirect. Available online: <https://www.sciencedirect.com/science/article/pii/S0048969717327419> (accessed on 13 May 2022).
7. Fernández-Delgado, M.; Cernadas, E.; Barro, S.; Amorim, D. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* **2014**, *15*, 3133–3181.
8. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
9. Chen, W.; Li, Y.; Xue, W.; Shahabi, H.; Li, S.; Hong, H.; Wang, X.; Bian, H.; Zhang, S.; Pradhan, B.; et al. Modeling Flood Susceptibility Using Data-Driven Approaches of Naïve Bayes Tree, Alternating Decision Tree, and Random Forest Methods. *Sci. Total Environ.* **2020**, *701*, 134979. [CrossRef] [PubMed]
10. Lee, S.; Kim, J.-C.; Jung, H.-S.; Lee, M.J.; Lee, S. Spatial Prediction of Flood Susceptibility Using Random-Forest and Boosted-Tree Models in Seoul Metropolitan City, Korea. *Null* **2017**, *8*, 1185–1203. [CrossRef]
11. Vafakhah, M.; Mohammad Hasani Loor, S.; Pourghasemi, H.; Katebikord, A. Comparing Performance of Random Forest and Adaptive Neuro-Fuzzy Inference System Data Mining Models for Flood Susceptibility Mapping. *Arab. J. Geosci.* **2020**, *13*, 417. [CrossRef]
12. Youssef, A.M.; Pourghasemi, H.R.; El-Haddad, B.A. *Advanced Machine Learning Algorithms for Flood Susceptibility Modeling-Comparison of Their Performance; Safaga-Ras Gharib Area: Red Sea, Egypt*, 2022.
13. Ardabili, S.; Mosavi, A.; Várkonyi-Kóczy, A.R. Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. In Proceedings of the Engineering for Sustainable Future; Várkonyi-Kóczy, A.R., Ed.; Springer International Publishing: Cham, Switzerland, 2020; pp. 215–227.
14. Kazienko, P.; Lughofer, E.; Trawiński, B. Hybrid and Ensemble Methods in Machine Learning J. UCS Special Issue. *J. Univers. Comput. Sci.* **2013**, *19*, 457–461.
15. Sikora, R.; Al-Laymoun, O. A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms. Available online: <https://www.igi-global.com/chapter/a-modified-stacking-ensemble-machine-learning-algorithm-using-genetic-algorithms/www.igi-global.com/chapter/a-modified-stacking-ensemble-machine-learning-algorithm-using-genetic-algorithms/122748> (accessed on 21 April 2022).
16. Agriculture and Agri-Food Canada. Terrestrial Ecoregions of Canada-Open Government Portal. Available online: <https://open.canada.ca/data/en/dataset/ade80d26-61f5-439e-8966-73b352811fe6> (accessed on 9 May 2022).
17. Natural Resources Canada. Physiographic Regions-Open Government Portal. Available online: <https://open.canada.ca/data/en/dataset/028dd58d-320c-53fb-b5bc-8188fd5d5edf> (accessed on 9 May 2022).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.