Flood Driver Classification and Prediction in Canadian Watersheds

Mahkame Salimi Moghadam

December 2024

Abstract: Flooding is a frequent and impactful natural hazard in Canada, driven by diverse factors such as snowmelt, intense rainfall, and ice jams. Understanding the regional variations and underlying drivers of floods is essential for improving predictive capabilities and adaptive management strategies, particularly in the context of climate change. This study classifies 1,339 Canadian watersheds, characterized by 43 environmental and hydrogeomorphological attributes, into distinct flood driver categories. Utilizing advanced neural network models, the project aims to predict flood drivers in ungauged basins with high accuracy. The methodology involves statistical feature normalization, dimensionality reduction through Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), mutual information, random forest models, and machine learning-based classification. This approach offers a framework for region-specific flood prediction, aiding in risk mitigation and resource allocation.

Introduction and Background

Floods are among the most frequent and impactful natural hazards in Canada, driven by mechanisms such as snowmelt, intense rainfall, and ice jams. These mechanisms vary across regions due to environmental and hydrogeomorphological factors. Effective flood prediction requires understanding these drivers and their interaction with watershed characteristics. Climate change complicates flood prediction by introducing greater variability in precipitation, snowmelt timing, and extreme weather, making traditional models that rely on historical averages insufficients [1, 2].

Advanced data-driven methods, particularly machine learning, have emerged as powerful tools for analyzing the complex, high-dimensional data associated with floods. These methods classify watersheds into distinct flood driver categories and predict drivers in ungauged basins, providing insights for targeted flood risk management [3, 4].

The dataset utilized for this analysis includes 1,339 watersheds, each characterized by 43 environmental and hydrogeomorphological features, including topographical, hydrological, vegetation, shape, and climatic attributes. The dataset is provided by Natural Resources Canada (NRCan) and is classified as confidential. As a result, I submitted a mimic of the dataset within 100 data points, added some noise to all the values, and labelled the features by their categories' names.

To prepare the dataset for analysis, several preprocessing steps are applied, including feature selection techniques such as random forest importance, Recursive Feature Elimination (RFE), mutual information and correlation analysis to retain relevant features while reducing dimensionality and improving computational efficiency [5].

Flood drivers are not always mutually exclusive; a single watershed can experience multiple types of flooding mechanisms. This necessitates using multi-label classification techniques, where each instance can belong to multiple categories simultaneously. Neural networks are particularly well-suited for this task due to their ability to model complex, non-linear relationships within the data. Multi-label neural networks, leveraging binary cross-entropy loss and sigmoid activation, predict overlapping flood drivers accurately, even for ungauged basins [4]

Methods

This study systematically classifies Canadian watersheds by their primary flood drivers and predicts these drivers for ungauged basins. The methodology integrates data preparation, feature selection, dimensionality reduction, and machine learning-based classification.

1. Data Preparation: A dataset of 1,339 watersheds was preprocessed to address missing values and ensure completeness. Statistical normalization was applied to scale features, improving model convergence.

2. Feature Selection: Four methods were used to identify important features: random forest importance ranked features by their predictive contribution, RFE iteratively removed less significant features, mutual information measured feature dependency on targets, and the correlation matrix removed highly correlated features to reduce redundancy.

3. Dimensionality Reduction: PCA reduced the dataset's complexity by transforming the features into orthogonal components, retaining the most informative attributes for classification while minimizing data loss. PCA was applied to reduce data complexity while retaining 90% of the variance.

4. Model Training and Evaluation: A neural network designed for multi-label classification predicted overlapping flood drivers. I use the Top 20 features from the feature selection techniques to train my neural network model. The model used a sigmoid activation function for independent driver predictions and binary cross-entropy loss to handle multi-label outputs. The architecture included input and hidden layers with ReLU activation and an optimized output layer for each driver type. Hyperparameters were fine-tuned using grid search, and training employed the Adam optimizer for efficiency and adaptability.

5. Implementation Tools: Python libraries, including NumPy, Pandas, and Scikit-learn, were used for preprocessing, feature selection, and dimensionality reduction, while TensorFlow/Keras supported neural network design and training. Visualizations were generated with Matplotlib and Seaborn to present clear, interpretable results.

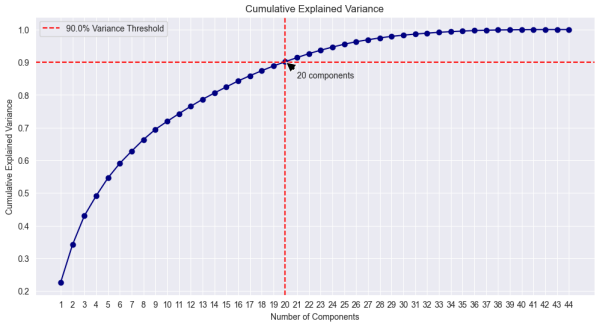
Results and Tests

This study analyzed feature contributions, reduced data dimensionality, and evaluated the performance of a multi-label neural network for flood driver classification. The findings are summarized as follows:

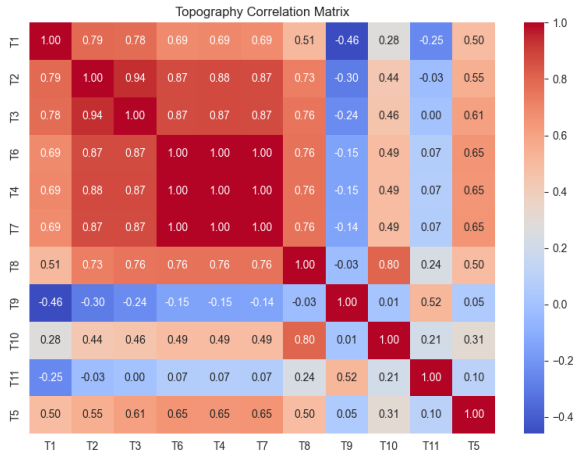
1. Dimensionality Reduction and Feature Importance:

PCA reduced the 43 features to 20 principal components, retaining 90% of the variance (Fig. 1a). This ensured computational efficiency without significant information loss.

Correlation analysis revealed redundancy among topographical features (`T1`, `T2`, `T3`), guiding feature removal, while unique features (`T10`, `T11`) provided non-redundant information (Fig. 1b).



(a) Cumulative Explained Variance.



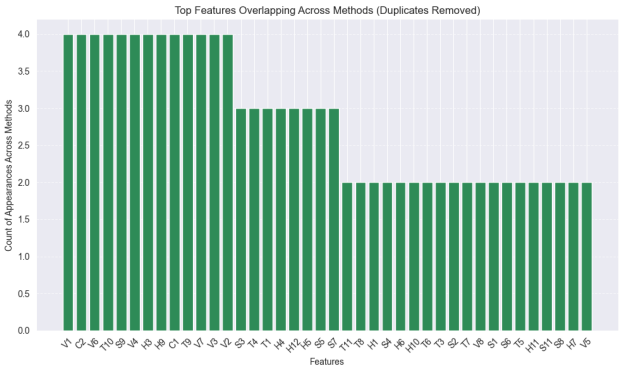
(b) Topography Feature Correlation Matrix.

Figure 1: (a) PCA components retaining 90% variance and (b) Correlation matrix for topographical features, highlighting strong positive correlations.

2. Feature Selection Analysis

Feature selection was performed using multiple techniques: RFE, random forest importance, PCA, and mutual information. The overlap of selected features across methods was analyzed to identify the most critical attributes. PCA identified significant contributors like `V7`, `T9`, and `H3`, while

(a) Feature Selection Overlap Heatmap.



(b) Count of Features Across Methods.

Figure 2: Feature selection results using RFE, random forest importance, PCA, and mutual information: (a) Heatmap showing selected feature overlaps, (b) bar plot of feature appearances across methods

mutual information ranked `V6`, `V4`, and `C2` as highly relevant (Fig. 2a). Features such as `C2`, `T9`, `C1`, `S9`, and `V3` (Fig. 2b) consistently emerged as the most significant contributors.

The overlap of selected features across multiple methods (RFE, PCA + mutual information, and random forest importance) is visualized in Fig. 3. This highlights both shared and unique features identified by the methods.

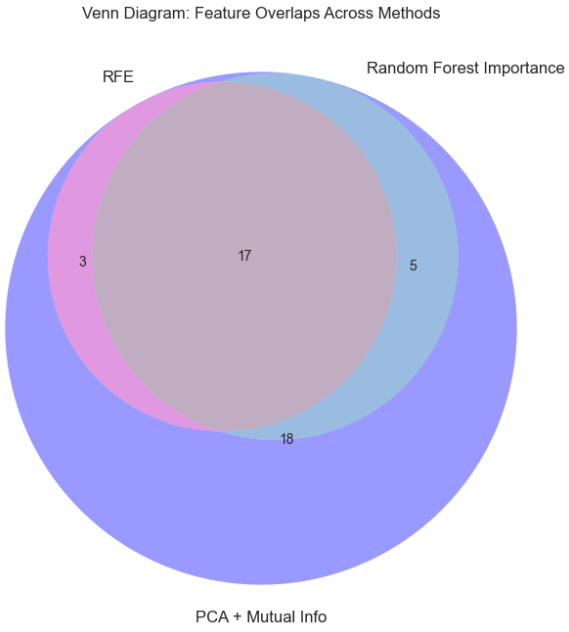


Figure 3: Venn Diagram of Feature Overlaps Across RFE, PCA + mutual information, and random forest importance Methods. The shared features (17) demonstrate consistency across techniques.

These features were retained in the final neural network model for improved predictive performance.

Neural Network Training and Evaluation: The neural network trained on the reduced feature set achieved stable convergence. Training and validation loss decreased consistently over 100 epochs (Fig. 4 left), with accuracy stabilizing around 40% after 40 epochs (Fig. 4 right). This demonstrates the model's ability to classify overlapping flood drivers in complex datasets effectively. The performance of the regularized neural network was evaluated over 100 epochs. Figure 4 shows the training and validation loss alongside accuracy.

Conclusions

This study aimed to classify Canadian watersheds based on their primary flood drivers and predict flood driver types using advanced machine learning techniques. The analysis incorporated dimensionality reduction, feature selection, and a multi-label neural network classifier.

PCA reduced the 43 original features to 20 components, capturing 90% of the total variance and ensuring computational efficiency without compromising critical information. Feature contri-

Figure 4: Training and Validation Loss (left) and Accuracy (right) for the Regularized Neural Network over 100 Epochs.

butions were assessed using PCA, mutual information, random forest importance, and RFE, with features such as `C2`, `T9`, `C1`, `S9`, and `V3` consistently identified as influential in determining flood drivers.

Pairwise correlation analysis revealed redundancies among topographical features, such as strong positive correlations between `T1`, `T2`, and `T3`, guiding the removal of redundant attributes to improve model efficiency.

The multi-label neural network achieved stable convergence over 100 epochs, with testing and validation accuracy stabilizing around 40%. Despite these results, further improvements are needed to enhance the model's efficiency and reliability for practical use.

Overall, the study highlights the importance of combining data-driven techniques for feature selection and dimensionality reduction with machine learning to address complex flood prediction problems. The identified features provide valuable insights into the environmental and hydrogeological characteristics influencing flood behavior across Canadian watersheds.

Future work can focus on refining the neural network architecture, such as ensemble models and attention mechanisms, incorporating techniques like hyperparameter optimization, exploring additional machine learning algorithms, and integrating spatiotemporal data to enhance prediction accuracy further. These advancements can support adaptive flood risk management strategies, especially in the context of climate change and its impact on flood dynamics.

For access to the project repository, code and code, visit the GitHub page: GitHub Repository.

References

[1] J. Singh, S. Ghosh, S. P. Simonovic, and S. Karmakar. Identification of flood seasonality and drivers across canada. *Hydrological Processes*, 35(10):e14398, 2021.

[2] Natural Resources Canada. Case studies on climate change in floodplain mapping. https://natural-resources.canada.ca/science-and-data/science-and-research/natural-hazards/flood-mapping/federal-flood-mapping-guidelines-series/case-studies-on-climate-change-floodplain-mapping/25468, 2018.

[3] L. Stein, M. P. Clark, W. J. M. Knoben, F. Pianosi, and R. A. Woods. How do climate and catchment attributes influence flood generating processes? a large-sample study for 671 catchments across the contiguous usa. *Water Resources Research*, 57(4):e2020WR028300, 2021.

[4] S. Sinsomboonthong. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *International Journal of Mathematics and Mathematical Sciences*, 2022(1):3584406, 2022.

[5] H. McGrath and P. N. Gohl. Prediction and classification of flood susceptibility based on historic record in a large, diverse, and data sparse country. *Environmental Sciences Proceedings*, 25(1), 2023.