ICFM 10
Adapting to Global Change:
Innovative Approaches to Flood Management and Resilience
May 20-22, 2026 | Western University, London, Ontario, Canada

PHLASH: A Physics-Informed Hybrid Learning Framework for Enhanced Short-Duration Flash Flood Forecasting

Farrukh A. Chishtie[1,2], Rana U. Ali[1], Abdolreza Bahremand[3], Mujtaba Hassan[4], John J. Clague[5]

[1]Peaceful Society, Science and Innovation Foundation, Vancouver BC, Canada

[2]Department of Occupational Science and Occupational Therapy, University of British Columbia, Vancouver BC, Canada

[3]Department of Watershed Management, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran

[4]Department of Space Science, Institute of Space Technology, Islamabad, Pakistan

[5]Centre for Natural Hazard Research, Simon Fraser University, Burnaby BC, Canada

E-mail: fachisht@uwo.ca

ABSTRACT

We introduce PHLASH (PHysics-informed Hybrid Learning for Accelerated Short-duration Hazards), an open framework designed to integrate short-duration, high-intensity rainfall data with terrain and soil factors for both flash flood susceptibility mapping and near-real-time forecasting. Using Google Earth Engine (GEE), we fuse hourly precipitation accumulations, meteorological data, MERIT DEM-derived terrain metrics such as HAND (Height Above Nearest Drainage) and slope, MODIS-based landcover/NDVI products, and daily soil moisture fields. The resulting unified dataset assigns pixel-level flood labels based on short-duration rainfall and local saturation thresholds. In trials conducted in Nova Scotia, Canada, we evaluated multiple machine learning algorithms with comprehensive metrics appropriate for imbalanced classification. We demonstrate that conventional metrics - overall accuracy and AUC - are misleading for rare event prediction, with models achieving 94% accuracy while missing 40% of flood events. An Artificial Neural Network with distribution-informed features achieves 90% recall and 90.6% balanced accuracy. SHAP analysis reveals that the intensity-duration product dominates predictions, validating fundamental hydrological understanding. PHLASH incorporates physics-based constraints including infiltration capacity and runoff routing directly into the learning pipeline. Although initially tailored for Nova Scotia, PHLASH is readily adaptable to other regions worldwide, delivering a robust, scalable early warning system that balances machine learning flexibility with physics-informed constraints.

KEYWORDS: PHLASH, Flash floods, Machine learning, Physics-informed learning, Google Earth Engine, Class imbalance, SHAP interpretability

1. INTRODUCTION

Climate change has increased the frequency and intensity of extreme weather events across all inhabited continents (IPCC, 2021). Among climate-related hazards, flash floods stand out as one of the most dangerous phenomena, distinguished by their rapid onset, localized intensity, and minimal warning time available for protective action (Jonkman, 2005). Addressing this challenge requires frameworks that integrate multi-source Earth observations with machine learning while incorporating physics-based constraints. We introduce PHLASH (PHysics-informed Hybrid Learning for Accelerated Short-duration Hazards), an open framework designed to fuse short-duration, high-intensity rainfall data with terrain and soil factors for both susceptibility mapping and near-real-time forecasting. Between 2000 and 2019, floods accounted for 44% of all disaster events, affecting over 1.65 billion people worldwide (CRED, 2020).

Recent catastrophic events underscore flash floods' devastating potential. On July 21-22, 2023, an atmospheric river delivered over 250 mm of rainfall to parts of Nova Scotia in less than 24 hours, killing four people and causing over $257 million in insured damages (IBC, 2024). The recurrence of fatal flash flooding in July 2024 demonstrates the persistent and escalating nature of this hazard.

Applying machine learning to flash flood prediction faces a fundamental challenge that is pervasively inadequately addressed: extreme class imbalance. Flash floods are rare events - occurring on perhaps 1-10% of days at any given location. In our dataset, flash flood events constitute 9% of observations. Yet this seemingly modest imbalance creates a critical evaluation problem. A model that predicts "no flood" for every instance achieves 91% accuracy while being completely useless for flood warning.

The flood prediction literature frequently reports AUC values exceeding 0.90 as evidence of model success (Band et al., 2020; He et al., 2025). However, examination of class-specific metrics often reveals a troubling pattern: high AUC coexists with inadequate minority class detection. A flood warning system that achieves 95% AUC but detects only 60% of actual floods provides a false sense of security.

The PHLASH framework incorporates distribution theory-informed feature generation through Weibull Extreme Value Theory analysis, building on the Distribution-Informed Graph Neural Network framework (Chishtie et al., 2024). Rather than manually constructing features through intuitive transformations, we derive features systematically from rigorous statistical characterization of the data-generating process. Our objectives are: (1) demonstrate the failure of conventional metrics, (2) introduce physics-informed feature generation using GEE, (3) evaluate multiple ML architectures with comprehensive metrics, and (4) provide interpretability through SHAP analysis.

2. MATERIALS AND METHODS

2.1 Study Area and Data

Nova Scotia, Canada, provides an ideal setting for flash flood susceptibility analysis (Figure 1). The province's complex terrain - ranging from coastal lowlands to interior highlands exceeding 270 m elevation - creates diverse hydrological conditions. The PHLASH framework uses Google Earth Engine (GEE) to fuse hourly precipitation accumulations covering 1-12 hour windows, meteorological data from ERA5-Land, MERIT DEM-derived terrain metrics including HAND (Height Above Nearest Drainage) and slope, MODIS-based landcover/NDVI products, and daily soil moisture fields. Flash flood susceptibility events were identified using Environment and Climate Change Canada (ECCC) operational warning thresholds combined with soil saturation criteria.

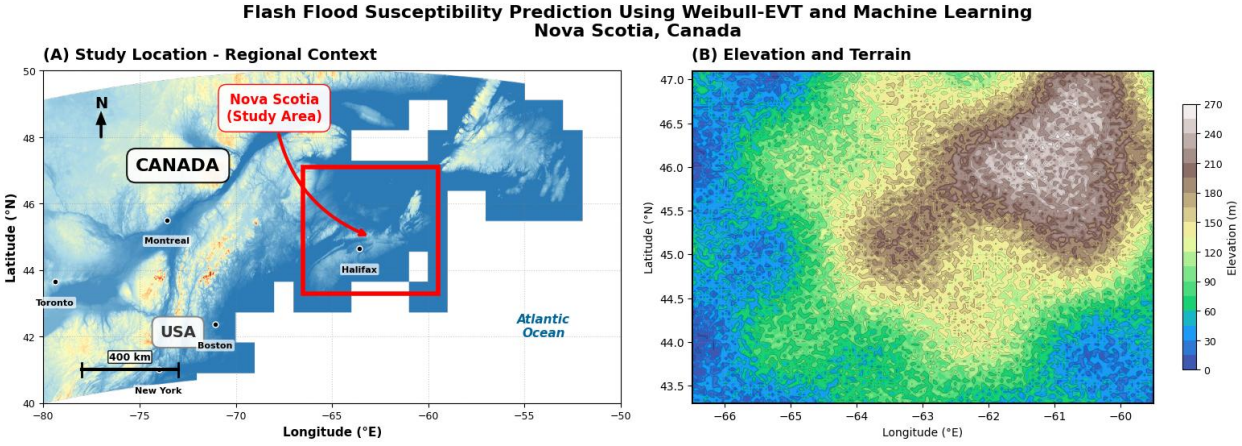

Figure 1: Study area overview for flash flood susceptibility prediction in Nova Scotia, Canada. (A) Regional context showing Nova Scotia's location in Atlantic Canada. (B) Digital elevation model revealing terrain variability across the study region, with elevations ranging from sea level through coastal plains to interior highlands exceeding 270 m.

The final dataset contains 1,117 samples: 100 flash flood events (9.0%) and 1,017 non-flood observations (91.0%). We constructed 28 baseline features comprising static terrain characteristics (slope, flow accumulation, Topographic Wetness Index, HAND) and dynamic meteorological variables from ERA5-Land (precipitation, soil moisture, runoff).

2.2 Physics-Informed Feature Generation

PHLASH incorporates physics-based constraints directly into the learning pipeline. Extreme Value Theory provides the mathematical framework for characterizing distribution tails where rare events occur (Coles, 2001). Preliminary analysis revealed bounded tail behavior in Nova Scotia precipitation data, indicating the Weibull distribution is appropriate. The two-parameter Weibull distribution is defined by $F(x; k, \lambda) = 1 - \exp[-(x/\lambda)^k]$, where k is the shape parameter and lambda is the scale parameter representing the characteristic extreme value.

From pooled precipitation data, we computed Weibull shape k = 2.71 and scale lambda = 279.87 mm. The scale parameter represents the approximate physical upper bound for daily precipitation - consistent with maximum observed precipitation during intense extratropical cyclones. From the fitted distribution, we derived 24 features including exceedance probabilities, return periods, threshold-relative metrics, and physical process interactions such as intensity-duration product and rain-on-saturated-soil.

2.3 Machine Learning Models

We evaluated three architectures: Random Forest (100 trees, depth 12, balanced sample weights), Support Vector Machine (RBF kernel, C = 0.1, balanced class weights), and Artificial Neural Network (128-64-32-1 architecture with focal loss). Data were split 80%/20% with stratification. Feature selection using mutual information identified the top 20 features, balancing model complexity against generalization.

2.4 Evaluation Metrics

For imbalanced classification, we report balanced accuracy (arithmetic mean of class-specific accuracies), precision, recall, F1-score, Critical Success Index, and bias score. These metrics directly characterize minority class performance that accuracy and AUC obscure. SHAP analysis was applied to quantify feature contributions to individual predictions.

3. RESULTS

3.1 Demonstration of Metric Failure

Table 1 presents comprehensive performance across all model configurations, revealing how conventional metrics mask operational failures.

Table 1: Comprehensive model performance on test set (224 samples: 20 floods, 204 non-floods)

Model	Acc.	Bal.Acc.	Prec.	Recall	F1	AUC
RF Base	0.942	0.788	0.706	0.600	0.649	0.975
RF Selected	0.929	0.826	0.583	0.700	0.636	0.967
SVM Base	0.786	0.882	0.294	1.000	0.455	0.949
SVM Selected	0.826	0.904	0.339	1.000	0.506	0.946
ANN Weibull	0.911	0.816	0.500	0.700	0.583	0.936
ANN Selected	**0.911**	**0.906**	**0.500**	**0.900**	**0.643**	**0.950**

The results demonstrate that RF Base achieves the highest accuracy (94.2%) and AUC (0.975) while detecting only 60% of flood events. Conversely, SVM models achieve lower accuracy (78.6-82.6%) but

detect all 20 flood events. The model with highest accuracy misses 8 flood events; the model with lowest accuracy misses none. AUC rankings are inversely related to operational flood detection performance.

3.2 Best-Performing Model

The ANN with selected features achieves the best balance: 90% recall (detecting 18 of 20 floods), 90.6% balanced accuracy, and generating approximately equal numbers of true positives and false positives. This performance represents viable operational deployment, detecting the vast majority of flood events while maintaining acceptable precision.

3.3 SHAP Interpretability

SHAP analysis reveals that the intensity-duration product dominates predictions with mean |SHAP| = 0.127 - more than triple the second-ranked feature (Figure 2). This compound variable captures total rainfall energy delivered to the landscape, validating fundamental hydrological understanding. Four of the top seven SHAP-ranked features are distribution-informed (intensity-duration product, runoff ratio, distance from bound, precipitation excess), demonstrating that physics-informed features provide predictive value beyond conventional meteorological variables.

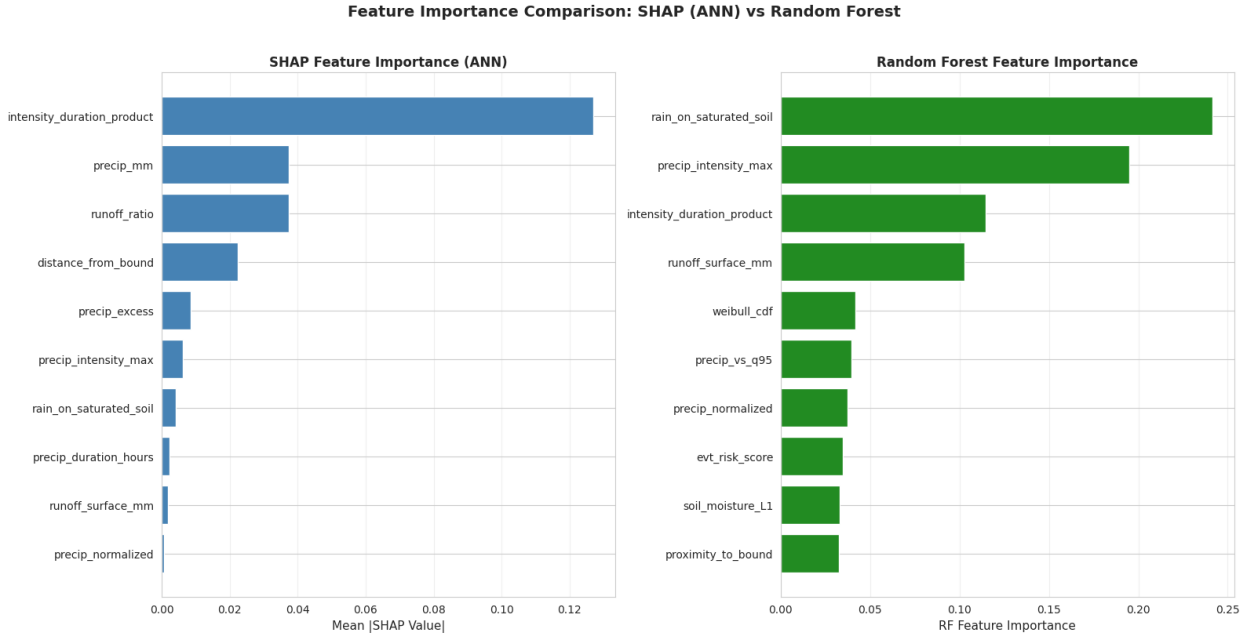

Figure 2: Feature importance comparison between SHAP analysis of the ANN model (left) and Random Forest permutation importance (right). The intensity-duration product dominates ANN predictions with mean |SHAP| = 0.127. Both algorithms identify physical process features as dominant predictors, validating the PHLASH framework.

4. DISCUSSION

Our results provide compelling empirical evidence for the failure of conventional metrics in imbalanced classification. Overall accuracy's failure stems from its domination by majority class performance. With 91% non-flood observations, accuracy = 0.91 x Specificity + 0.09 x Recall - systematically penalizing flood detection. AUC measures ranking quality rather than operational performance at any specific threshold, explaining why the model with highest AUC (0.975) achieves lowest recall (0.60).

The PHLASH framework demonstrates both theoretical and practical value by incorporating physics-based constraints directly into the learning pipeline. By fitting the Weibull distribution to precipitation

extremes, we obtain principled features: the threshold u = 125.15 mm formally defines "extreme" precipitation; the scale parameter lambda = 279.87 mm provides the physical upper bound; exceedance probabilities give precise likelihood estimates. These features are interpretable characterizations of extreme behavior rather than intuitive approximations.

For practitioners, we recommend: (1) report balanced accuracy, precision, recall, and confusion matrices for all rare event prediction models; (2) do not rely on accuracy or AUC as primary evaluation measures; (3) incorporate physics-informed features including HAND, infiltration capacity, and runoff routing; (4) use physical process interactions that combine distribution-derived normalization with hydrological understanding.

5. CONCLUSION

This study introduces PHLASH (PHysics-informed Hybrid Learning for Accelerated Short-duration Hazards), advancing flash flood susceptibility prediction through three contributions. First, we demonstrate empirically that both accuracy and AUC are fundamentally misleading for imbalanced classification - models achieving 94% accuracy and AUC exceeding 0.97 simultaneously miss 40% of flood events. Second, we integrate physics-based constraints including terrain metrics, infiltration capacity, and runoff routing directly into the learning pipeline using Google Earth Engine. Third, SHAP analysis provides interpretability essential for operational deployment, with the intensity-duration product emerging as the dominant predictor.

Our best-performing model achieves 90% recall and 91% balanced accuracy, detecting 18 of 20 flood events. Although initially tailored for Nova Scotia's short-duration flood hazards, PHLASH is readily adaptable to other regions worldwide, provided suitable hourly precipitation, DEM, and landcover datasets are available within GEE. The framework delivers a robust, scalable early warning system that balances machine learning flexibility with physics-informed constraints, addressing the critical demand for accurate, timely, and interpretable flood hazard mapping.

6. ACKNOWLEDGEMENTS

F.A.C. acknowledges funding from Vancouver Foundation via their Recovery and Resilience grant awarded to Peaceful Society, Science and Innovation Foundation. ERA5-Land data were accessed through Google Earth Engine.

7. REFERENCES

Band S.S., Janizadeh S., Chandra Pal S., Saha A., Chakrabortty R., Shokri M. and Mosavi A. (2020). Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. Remote Sensing, 12(21), 3568.

Chishtie F.A., Chishtie F., Mushtaq H., Lam A.C.L. and Hsieh W.W. (2024). Advancing heatwave forecasting via distribution informed-graph neural networks: Integrating extreme value theory with GNNs. arXiv preprint arXiv:2411.13496.

Coles S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London.

CRED (2020). EM-DAT: The International Disaster Database. Centre for Research on the Epidemiology of Disasters, Brussels.

He F., Liu S., Mo X. and Wang Z. (2025). Interpretable flash flood susceptibility mapping in Yarlung Tsangpo River Basin using H2O Auto-ML. Scientific Reports, 15, 1702.

IBC (2024). Anniversary of Nova Scotia's deadly flash flood provides sombre reminder of the impacts of severe weather. Insurance Bureau of Canada.

IPCC (2021). Climate Change 2021: The Physical Science Basis. Cambridge University Press, Cambridge.

Jonkman S.N. (2005). Global perspectives on loss of human life caused by floods. Natural Hazards, 34(2), 151-175.