

XGBoost-Based Optimal Transport Dataset Distance-Guided Temporal Domain Transfer for Flood Prediction Under Concept Drift

Sharifah Naluswata

snaluswata18@gmail.com

Victoria University

Jackson Kisuule

Victoria University

David Kakeeto

Victoria University

Ali Najib

Victoria University

Anthony Bua

Victoria University

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11114447/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

XGBoost-Based Optimal Transport Dataset Distance-Guided Temporal Domain Transfer for Flood Prediction Under Concept Drift

Naluswata Sharifah^{1*†}, Kisuule Jackson^{1*†}, Kakeeto David^{1†}, Najib Ali^{1†},
Bua Anthony^{1†}

¹Faculty of science and technology, Victoria University, Kampala

**Corresponding authors: snaluswata18@gmail.com,
jkisuule89@gmail.com*

Abstract

Floods remain a major natural disaster that causes loss of life, infrastructure damage, and economic losses worldwide . Although machine learning has improved flood prediction, most existing models cannot effectively adapt to changing environmental conditions caused by concept drift. Therefore, this study sought to develop an adaptive model capable of maintaining accurate and reliable flood predictions under evolving hydrological conditions. An XGBoost-Based Optimal Transport Dataset Distance (OTDD)-Guided Temporal Domain Transfer Learning Model was developed using a publicly available flood prediction dataset obtained from Kaggle, integrated Extreme Gradient Boosting as the primary prediction model, Adaptive Windowing for automatic concept drift detection, and OTDD for measuring distributional differences between historical and incoming environmental data to facilitate temporal domain transfer learning. The developed model was subsequently evaluated using the performance metrics of accuracy, precision, recall, F1-score, ROC-AUC, Root Mean Squared Error (RMSE), OTDD alignment score, and computational execution time. The experimental evaluation demonstrated that XGBoost-OTDD Temporal Domain Transfer Model achieved superior predictive performance compared with the baseline models.

The model attained an accuracy of 88.7%, precision of 82.9%, recall of 91.6%, and an F1-score of 87.1%, outperforming both Logistic Regression and Static XGBoost across all evaluated metrics.

Keywords: Machine learning; Concept drift; Optimal Transport Dataset Distance; Flood prediction.

1 Introduction

Flooding is one of the most destructive natural hazards worldwide [1], causing substantial loss of life, displacement, infrastructure damage, and economic disruption. The challenge is particularly significant in developing regions where rapid urbanization, changing rainfall patterns, limited drainage infrastructure, and inadequate hydrological monitoring increase community vulnerability to flood events [2]. In Uganda and the wider East African region, recurrent floods and landslides continue to affect vulnerable communities, highlighting the need for reliable and timely flood prediction systems capable of supporting early warning and disaster preparedness [2].

Flood prediction has progressively evolved from conventional statistical and hydrodynamic approaches toward data-driven machine learning (ML) and deep learning techniques [3–5]. Machine learning models can learn complex nonlinear relationships among hydro-meteorological variables such as rainfall, river discharge, soil moisture, temperature, humidity, and atmospheric pressure, making them increasingly useful for flood forecasting [6–8]. Algorithms including Logistic Regression, Random Forest, Support Vector Machines, Artificial Neural Networks, and gradient boosting methods have consequently been investigated for flood prediction with promising results [9–12].

Among these techniques, Extreme Gradient Boosting (XGBoost) has demonstrated strong predictive capability [11] because of its regularization, computational efficiency, ability to model nonlinear interactions, and robustness on structured data. Recent flood prediction studies have reported competitive performance from XGBoost and related ensemble methods [13–16]. However, high predictive accuracy obtained under static experimental conditions does not necessarily guarantee reliable performance after deployment. Most conventional models assume that training and future observations follow approximately the same probability distribution [17, 18]. Hydrological environments violate this assumption because rainfall regimes, temperature, soil conditions, land use, river discharge, and other environmental processes change

over time [19].

These temporal changes can produce *concept drift*, where the statistical properties of incoming observations or the relationship between predictors and flood occurrence differ from those represented by the historical training data [20]. Under such conditions, static prediction models may experience decision-boundary degradation and increasing classification error [?]. Although complete model retraining can restore predictive performance, frequent retraining is computationally expensive and may discard useful knowledge learned from historically relevant hydrological periods. Adaptive mechanisms that determine when distributional change has occurred and which historical knowledge remains transferable are therefore required [21].

Concept-drift detection methods provide one approach for identifying changes in evolving data streams [22]. Adaptive Windowing (ADWIN), for example, automatically compares sub-windows of streaming observations and signals drift when statistically significant differences are detected [23, 24]. Detecting drift alone does not determine how knowledge should be transferred from historical data to a new hydrological regime [25]. Transfer learning and domain adaptation address this problem by exploiting knowledge acquired from a source domain to improve learning in a related target domain [26]. Their effectiveness, however, depends strongly on the similarity between source and target distributions; transferring knowledge from a poorly matched domain can result in negative transfer [26, 27].

Optimal Transport (OT) provides a principled mathematical framework for quantifying differences between probability distributions [28–30]. Wasserstein-based optimal transport measures preserve the geometry of distributions and have therefore become increasingly important in domain adaptation and distribution alignment [29, 31]. Optimal Transport Dataset Distance (OTDD) extends this principle to dataset-level comparison by incorporating feature and label information when measuring the distance between datasets [32]. This property makes OTDD particularly suitable for temporal flood prediction, where historical hydrological periods differ in their relevance to current environmental conditions.

Despite these advances, existing flood prediction studies predominantly optimize predictive performance under static data distributions [33]. Other studies such as [34] address concept-drift detection, domain adaptation, transfer learning, or distributional alignment independently. Consequently, limited

attention has been given to a unified flood prediction framework that detects temporal concept drift, quantitatively measures the similarity between historical and incoming hydrological domains, and selectively transfers relevant knowledge to an adaptive predictor. This limitation is especially important in non-stationary hydrological environments where indiscriminate use of historical observations may reduce prediction reliability.

To address this gap, this study develops an *XGBoost-Based Optimal Transport Dataset Distance (OTDD)-Guided Temporal Domain Transfer Model for Flood Prediction under Concept Drift*. The proposed framework integrates XGBoost as the primary flood classifier, ADWIN for automatic concept-drift detection, and OTDD for measuring temporal distributional divergence and guiding knowledge transfer between historical and incoming hydrological domains. Rather than repeatedly rebuilding the prediction model from scratch, the framework identifies distributionally relevant historical information and uses it to adapt learning when environmental conditions change.

2 Methods

2.1 Research Design and Experimental Framework

This study adopted an experimental machine learning design to develop and evaluate an XGBoost-Based Optimal Transport Dataset Distance Guided Temporal Domain Transfer Model for flood prediction under concept drift. The framework combined three principal components: Extreme Gradient Boosting (XGBoost) as the predictive learner, Adaptive Windowing (ADWIN) for concept drift detection, and OTDD for quantifying distributional divergence between historical and incoming temporal domains. The experimental procedure comprised data acquisition and preprocessing, chronological temporal partitioning, baseline model development, concept drift detection, OTDD-based domain comparison, temporal knowledge transfer, adaptive XGBoost prediction, and performance evaluation.

2.2 Dataset Description and Preprocessing

The proposed model was developed and evaluated using a publicly available flood prediction dataset obtained from Kaggle ¹ containing historical hydro-

¹<https://www.kaggle.com/datasets/farahfirdausa/flood-prediction-dataset>

logical, meteorological, and environmental observations associated with flood occurrence. The target variable represented the occurrence of floods and was used for supervised binary classification.

Prior to modelling, the dataset was subjected to an integrity audit to identify missing values, duplicate observations, inconsistent data types, and invalid records. Numerical variables were converted to appropriate numeric formats, while the date attribute was transformed into a standard date-time representation to preserve the temporal ordering of observations. Exploratory and descriptive analyses were subsequently performed to examine variable distributions, class frequencies, correlations, and potential distributional changes.

To prevent information leakage between historical and future observations, the cleaned dataset was ordered chronologically rather than randomly shuffled. The resulting observations were then organized into sequential temporal blocks representing evolving hydrological periods.

2.3 Development of the Model

The complete flood prediction dataset was represented as;

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N, \quad (2.1)$$

Where; $\mathcal{D}$ denoted the complete dataset, x_i represented the predictor variables, y_i denoted the flood occurrence label, and N was the total number of observations.

Each predictor vector was represented as;

$$x_i = [x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}]^T \quad (2.2)$$

Where; m denoted the number of predictor variables including rainfall, river discharge, humidity, temperature, pressure, wind speed, soil moisture, elevation, and other environmental characteristics. The study employed a 70% training and 30% testing split.

2.3.1 ADWIN

ADWIN was employed to detect concept drift within the continuous stream of environmental observations. Unlike conventional methods that use a fixed-size window, ADWIN automatically adjusts the window size according to the

statistical characteristics of the incoming data. This enabled the model to detect significant environmental changes while ignoring insignificant fluctuations caused by random noise. Adaptive window was represented by;

$$W = \{x_1, x_2, \dots, x_n\} \quad (2.3)$$

Where; n denoted the current window size.

The adaptive window was divided into two sub windows,

$$W = W_0 \cup W_1 \quad (2.4)$$

Where; W_0 contained historical observations, while W_1 contained the most recent observations.

The average values of the two windows were computed as

$$\mu_0 = \frac{1}{|W_0|} \sum_{x \in W_0} x \quad (2.5)$$

and

$$\mu_1 = \frac{1}{|W_1|} \sum_{x \in W_1} x \quad (2.6)$$

The difference between the two means was then calculated as;

$$\Delta\mu = |\mu_0 - \mu_1| \quad (2.7)$$

To determine whether the observed difference was statistically significant, ADWIN employed the Hoeffding Bound

$$\varepsilon = \sqrt{\frac{1}{2m} \ln \left(\frac{4}{\delta} \right)} \quad (2.8)$$

Where;

$$m = \frac{|W_0||W_1|}{|W_0| + |W_1|} \quad (2.9)$$

And δ denoted the confidence level selected by the researcher. Concept drift was declared whenever

$$|\mu_0 - \mu_1| > \varepsilon \quad (2.10)$$

If Equation 2.10 was satisfied, the historical window W_0 was discarded while the recent window W_1 became the new adaptive window.

The updated adaptive window therefore became

$$W = W_1 \quad (2.11)$$

This adaptive mechanism enabled the proposed flood prediction model to continuously respond to changing environmental conditions while avoiding unnecessary retraining whenever no significant drift exists.

2.3.2 OTDD

After ADWIN detects concept drift, OTDD measures the similarity between historical and new environmental data. The model transfers previous knowledge when similarity is high and updates itself when significant changes occur. The historical dataset (source domain) was represented as;

$$D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s} \quad (2.12)$$

Where; x_i^s represented the feature vector in the source domain, y_i^s represented the corresponding flood label, and n_s denoted the number of observations in the historical dataset.

Similarly, the newly arriving environmental dataset (target domain) was represented as

$$D_t = \{(x_j^t, y_j^t)\}_{j=1}^{n_t} \quad (2.13)$$

Where; n_t denoted the number of observations in the target dataset. The OTDD between the two datasets was defined as

$$OTDD(D_s, D_t) = \min_{\gamma} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \gamma_{ij} c(x_i^s, x_j^t) \quad (2.14)$$

Where; γ_{ij} denoted the transport plan, and $c(x_i^s, x_j^t)$ represented the transportation cost between the two observations. The transportation plan satisfied the following marginal constraints

$$\gamma \mathbf{1} = a$$

and

$$\gamma^T \mathbf{1} = b$$

Where; a and b represent the probability distributions of the source and target datasets respectively.

The transportation cost was computed using the Euclidean distance

$$c(x_i, x_j) = \|x_i - x_j\|_2^2 \quad (2.15)$$

The computed OTDD value was then compared with a predefined similarity threshold

$$OTDD(D_s, D_t) \leq \tau \quad (2.16)$$

Where; τ denote the transfer threshold selected during experimentation. If Equation 2.16 was satisfied, the source knowledge was transferred directly to the target domain without retraining. Otherwise, a new learning process was initiated to accommodate the observed environmental changes.

2.3.3 Temporal Domain Transfer

Temporal Domain Transfer reuses knowledge from historical flood data when OTDD identifies sufficient similarity with new environmental conditions. It adapts previously learned model parameters instead of fully retraining, reducing computational cost while improving generalization. The source domain feature representation was denoted as

$$F_s = \{f_1, f_2, \dots, f_m\} \quad (2.17)$$

while the adapted target feature representation became

$$F_t = F_s + \Delta \quad (2.18)$$

Where; Δ represent the temporal adaptation learned from newly arriving environmental observations. The adaptation function was expressed as

$$\Delta = \phi(D_t) - \phi(D_s) \quad (2.19)$$

Where; $\phi()$ represent the learned feature transformation.

The transferred prediction model was therefore represented as

$$M_t = M_s + \Delta M \quad (2.20)$$

Where; M_s denote the previously trained flood prediction model, and ΔM represent the model adjustment after concept drift.

2.3.4 XGBoost

The adapted dataset from concept drift detection and temporal domain transfer was used to train an XGBoost classifier for accurate flood prediction. Hyperparameter optimization and cross-validation improved generalization while reducing overfitting and prediction bias.

For each observation, the predicted flood occurrence was expressed as

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2.21)$$

Where; f_k represent the k^{th} regression tree, and K denote the total number of trees.

The XGBoost objective function was defined as

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2.22)$$

Where; $l()$ represents the prediction loss function, and $\Omega(f_k)$ is the regularization term controlling model complexity.

The regularization function was given by

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2.23)$$

Where; T represent the number of terminal nodes, w_j denote the weight associated with the j^{th} leaf, while γ and λ are regularization parameters that minimize model overfitting.

During each boosting iteration, the objective function was approximated using second-order Taylor expansion as

$$L^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (2.24)$$

Where; g_i and h_i represent the first-order and second-order derivatives of the loss function respectively.

The final flood prediction is obtained using the sigmoid transformation

$$P(\text{Flood}) = \frac{1}{1 + e^{-\hat{y}}} \quad (2.25)$$

Where; $P(\text{Flood})$ denote the probability of flood occurrence if

$$P(\text{Flood}) \geq 0.5 \quad (2.26)$$

the observation was classified as a flood event; otherwise, it was classified as a non-flood event.

2.4 Model Architecture

Fig. 1 presents the complete adaptive flood prediction architecture from data preprocessing to final prediction. ADWIN continuously monitors environmental data streams to detect concept drift and trigger model adaptation when necessary. OTDD measures distributional similarity between historical and incoming observations to guide temporal knowledge transfer. Based on similarity, the framework either adapts existing knowledge or incrementally updates the model for new environmental conditions. XGBoost then models nonlinear environmental relationships and generates binary flood or no-flood predictions. Performance is evaluated using Accuracy, Precision, Recall, F1-Score, AUC-ROC, and computational efficiency.

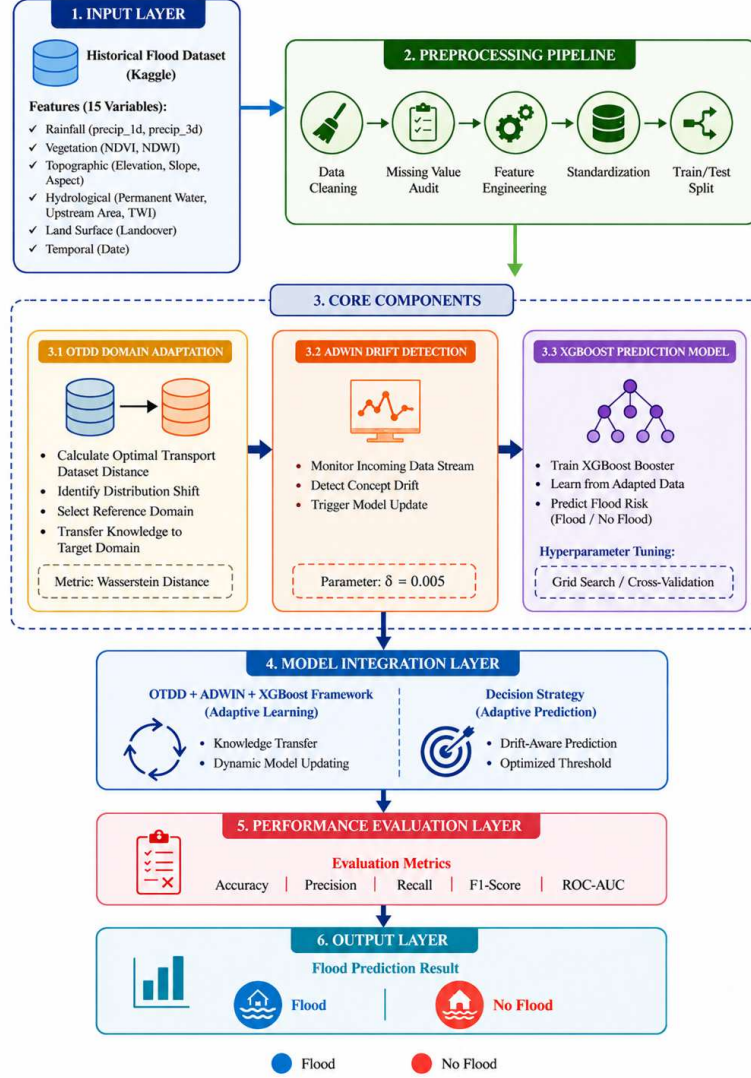


Fig. 1: XGBoost-Based OTDD-Guided Temporal Domain Transfer Model Architecture for Flood Prediction

2.5 Model Simulation

The experiments were implemented in Python using a computing environment equipped with an Intel Core i5 processor, 16 GB RAM, Windows 11,

and a 512 GB SSD. Pandas and NumPy were used for data manipulation and numerical processing, Scikit-learn for preprocessing and model evaluation, XGBoost for gradient-boosted classification, River for ADWIN-based concept drift detection, and the Python Optimal Transport (POT) library for optimal transport computations. Matplotlib and Seaborn were used for exploratory and performance visualization. Model development was conducted using Google Colab ².

3 Results and Discussion

3.1 Distribution of the Target Variable

The target variable was analyzed to determine the representation of flood and non-flood observations in the dataset. The target was binary-coded, where 0 represented no flood and 1 represented flood occurrence as seen in Table 1 . The results showed that the dataset was highly imbalanced. Non-flood observations accounted for approximately 95.48% of the dataset, whereas flood observations represented only 4.52%. This meant that the model encountered considerably more non-flood cases than actual flood cases during learning. Such an imbalance could cause a conventional classifier to favor the majority class and achieve high overall accuracy while failing to correctly identify the comparatively rare flood events.

Table 1: Distribution of the Target Class

Target class	Description	Frequency	Proportion
0	No flood	979,446	95.4811%
1	Flood	46,355	4.5189%
Total		1,025,801	100.0000%

3.2 Correlation analysis

The correlation analysis further showed that most predictors had weak linear relationships, suggesting limited multicollinearity. As shown in Fig. 2, the

²<https://colab.research.google.com/drive/1ikv067TWdZVedwjXEA31LEFq4q6BwPo#scrollTo=qW7pGgvQNbbw>

strongest relationship occurred between NDVI and NDWI ($r = 0.64$), while one-day and three-day precipitation exhibited a moderate positive relationship ($r = 0.49$). Elevation and TWI showed generally weak associations with the remaining predictors. These results indicate that the variables capture complementary dimensions of the environmental conditions associated with flood occurrence.

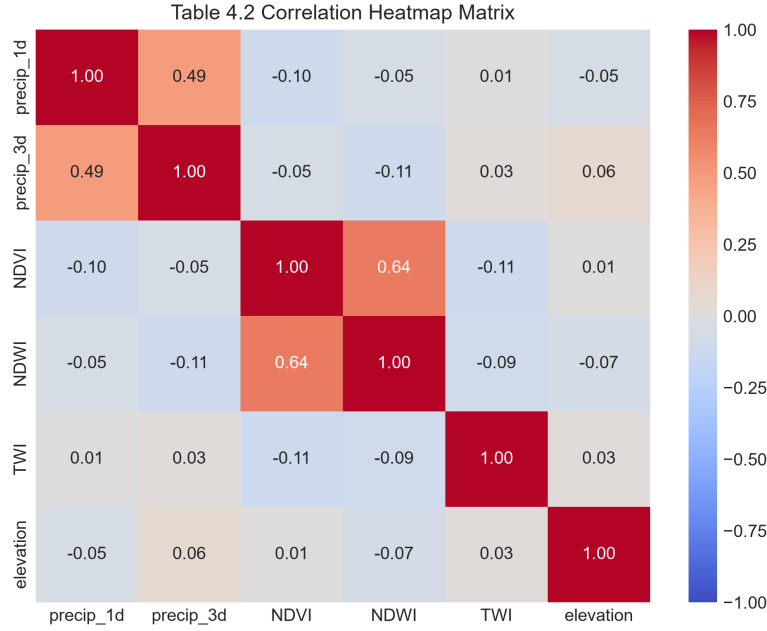


Fig. 2: Correlation matrix of selected flood prediction variables.

3.3 Concept Drift and OTDD-Based Distributional Alignment

Fig. 3 presents the temporal behavior of ADWIN and the OTDD alignment measure across six chronological hydrological eras. Eras 0–2 remained stable, with OTDD values at or close to zero and no corresponding drift alerts. This indicates that the early temporal blocks remained distributionally similar to the historical reference domain.

A substantial change emerged in Era 3, where the OTDD value increased sharply to approximately 73.18. The increase coincided with an ADWIN drift alert, providing evidence of a major change in the incoming environmental

distribution.

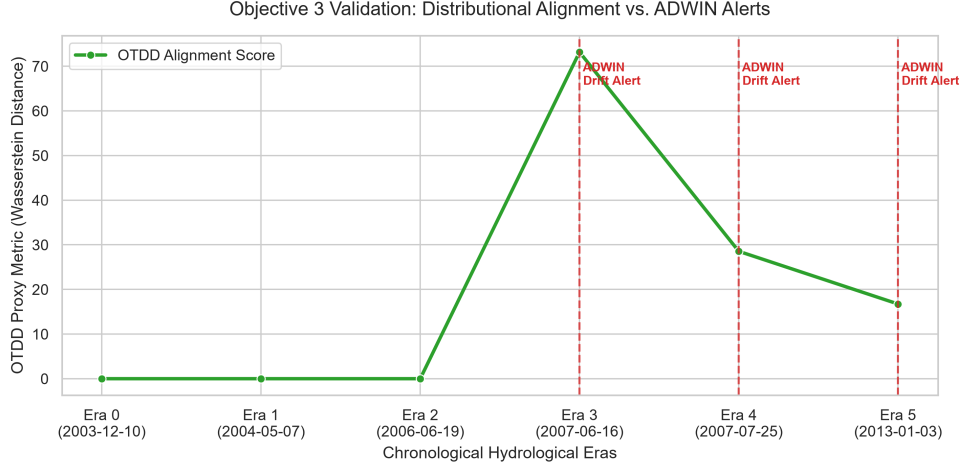


Fig. 3: Distributional alignment and ADWIN concept drift alerts across hydrological eras

Following adaptive updating, OTDD decreased to approximately 28.54 in Era 4 and 16.74 in Era 5. Although ADWIN continued to detect changes in these later eras, the progressive reduction in OTDD indicates decreasing distributional divergence after adaptation. The two mechanisms therefore provided complementary information: ADWIN identified when statistically significant change occurred, whereas OTDD quantified the magnitude of the temporal distribution shift.

This finding is particularly important for temporal flood prediction because it demonstrates that a model trained on historical observations cannot be assumed to remain representative as environmental conditions evolve. The OTDD-guided adaptation mechanism provides a means of responding to such changes by transferring knowledge from distributionally relevant historical domains.

3.4 Prediction Error and Computational Efficiency

The effect of concept drift on predictive performance is shown in Fig. 4. Prediction error remained negligible during the initial stable eras but increased sharply in Era 3, where RMSE approached 1.0. This deterioration occurred

simultaneously with the largest OTDD value and the first major ADWIN drift alert, demonstrating the adverse effect of distributional change on a model operating under a previously learned environmental regime.

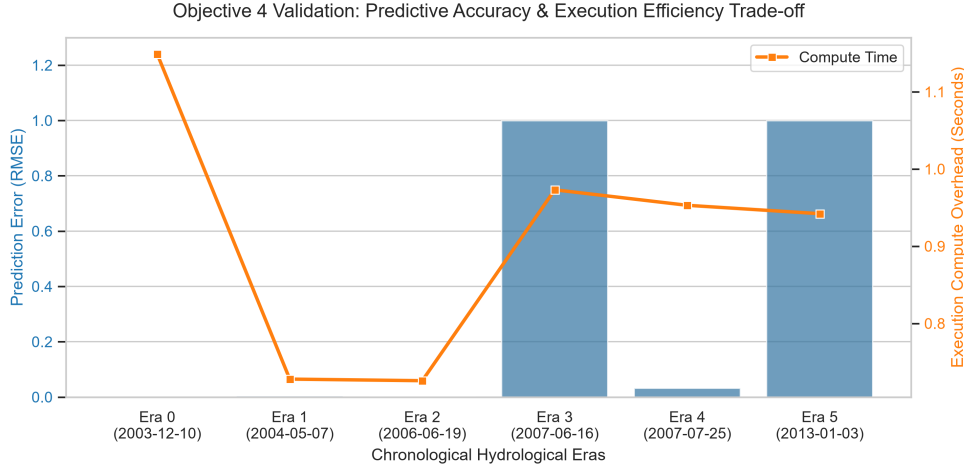


Fig. 4: Predictive error and execution-efficiency trade-off across hydrological eras.

After temporal adaptation, RMSE decreased sharply to approximately 0.03 in Era 4, indicating substantial recovery of predictive capability without complete model retraining. A slight subsequent increase in Era 5 suggests that the environmental process continued to evolve, reinforcing the need for continuous rather than one-time adaptation.

The computational cost of this adaptation remained relatively low. Processing time stayed below approximately one second for most hydrological eras, with only modest increases during drift-response operations. Thus, the improvement in distributional alignment and prediction error was achieved without substantial computational overhead, supporting the feasibility of the framework for near-real-time flood prediction.

3.5 Spatial Distribution of Flood Prediction Variables

The spatial distributions of precipitation, NDWI, elevation, and flood observations showed substantial geographical variability across the study area as show in Fig. 5. Areas with higher rainfall and moisture and lower elevation

exhibited conditions associated with greater flood susceptibility, highlighting the combined influence of multiple environmental predictors.

Spatial Distribution of Flood Prediction Variables

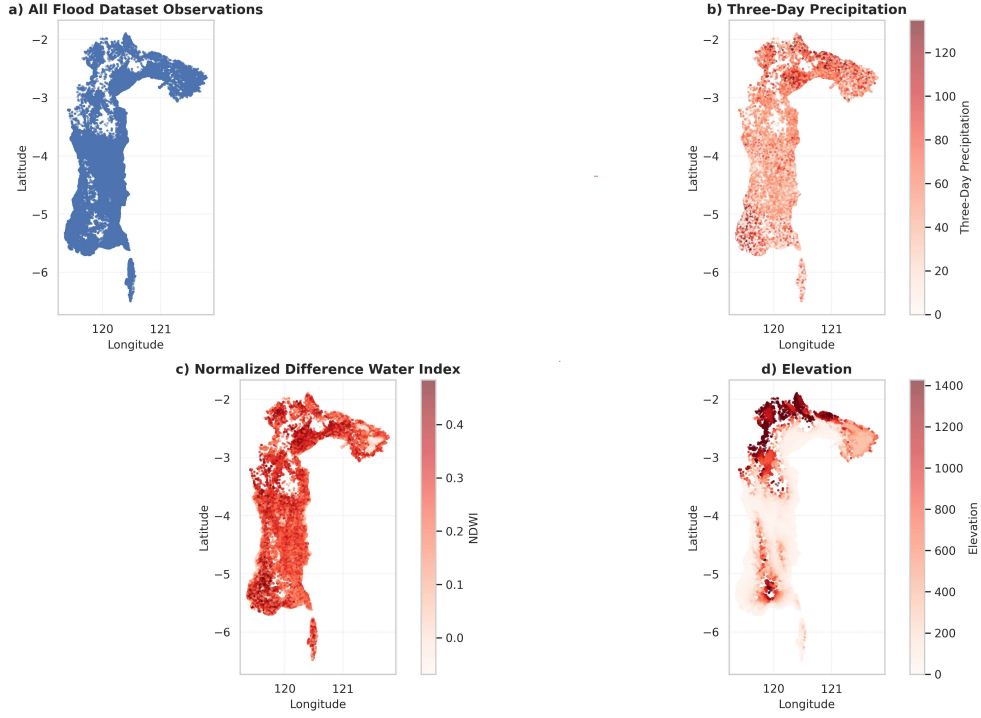


Fig. 5: Spatial distribution of flood prediction variables showing (a) observation locations, (b) three-day accumulated precipitation, (c) NDWI, and (d) elevation.

3.6 Discussion

The results demonstrate that temporal distribution change is an important source of predictive degradation in flood forecasting. The simultaneous rise in OTDD and RMSE during Era 3 provides direct experimental evidence that increasing divergence between historical and incoming environmental distributions adversely affects prediction performance. The subsequent decline in both OTDD and prediction error after adaptation indicates that selective temporal knowledge transfer can restore predictive capability.

These findings are consistent with previous studies showing that ensemble learners such as XGBoost are effective for modelling complex nonlinear flood relationships [35]. They also support concept-drift research showing that static machine learning models can lose predictive reliability when the underlying data-generating process changes [36]. However, the present study extends these approaches by integrating drift detection and distribution-distance measurement within the same adaptive prediction framework. The framework provides a mechanism for identifying *when* adaptation is required through ADWIN and estimating *how much* distributional change has occurred through OTDD. The decreasing OTDD values from approximately 73.18 in Era 3 to 28.54 and 16.74 in Eras 4 and 5, respectively, together with the reduction of RMSE to approximately 0.03 after adaptation, provide evidence that the proposed temporal transfer mechanism responded effectively to evolving environmental conditions.

The results indicate that integrating XGBoost, ADWIN, OTDD, and temporal domain transfer provides a more adaptive approach to flood prediction than relying exclusively on a static classifier. The low computational overhead further suggests that the model has potential for continuous flood-monitoring and early-warning applications where environmental distributions evolve over time.

4 Conclusion

This study developed an XGBoost-Based Optimal Transport Dataset Distance-Guided Temporal Domain Transfer Model for adaptive flood prediction under concept drift. The model integrates XGBoost for nonlinear flood prediction, ADWIN for automatic concept drift detection, and OTDD for quantifying distributional differences and guiding the transfer of relevant historical knowledge to changing hydrological conditions.

The experimental results demonstrated that temporal distribution shifts can substantially degrade the performance of static flood prediction models. A major concept drift was observed in Era 3, where the OTDD distance increased sharply and was accompanied by a corresponding increase in prediction error. Following adaptive temporal transfer, the distributional distance and prediction error decreased considerably in subsequent eras. These findings demonstrate that OTDD provides useful information on the magnitude

of environmental change, while ADWIN identifies when such change requires model adaptation. Importantly, this adaptation was achieved with relatively low computational overhead, with processing remaining below approximately one second for most temporal eras.

The model provides a unified mechanism for detecting concept drift, measuring temporal distributional divergence, and adapting an XGBoost predictor to evolving environmental conditions. The results indicate that combining OTDD-guided temporal transfer with concept drift detection can improve the robustness and adaptability of machine-learning-based flood prediction compared with relying solely on static learning. The framework therefore provides a promising foundation for continuous flood monitoring and data-driven early-warning applications in dynamic hydrological environments.

The study was evaluated using a publicly available historical flood dataset, and its performance under live operational data streams and different geographical regions remains to be established. Future research should therefore validate the model using real-time hydrological and meteorological sensor streams, evaluate its generalizability across multiple flood-prone regions, and incorporate additional spatial information such as remote-sensing and soil-moisture observations. Further investigation of spatio-temporal domain adaptation may also extend the model from temporal distribution shifts to simultaneous spatial and temporal environmental changes.

Abbreviations

ADWIN	Adaptive Windowing
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
OT	Optimal Transport
OTDD	Optimal Transport Dataset Distance
RMSE	Root Mean Squared Error
ROC-AUC	Receiver Operating Characteristic–Area Under the Curve
TWI	Topographic Wetness Index
XGBoost	Extreme Gradient Boosting

Acknowledgments

Not applicable.

Author Contributions

All co-authors equally contributed in the research and production of the paper.

Funding

This work is not supported by any organization or institution.

Data Availability Statement

The dataset generated during and/or analyzed during the current study is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval

Not applicable.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work

reported in this paper.

References

- [1] F. J. Glago, “Flood disaster hazards; causes, impacts and management: a state-of-the-art review,” *Natural hazards-impacts, adjustments and resilience*, 2021.
- [2] R. Twinomuhangi, H. Sseviiri, S. H. Mayhew, P. R. Group, and M. K. Sahani, “Examining the effects of climate and environmental hazards on vector and water-borne diseases in eastern uganda,” *Discover Environment*, vol. 4, no. 1, p. 104, 2026.
- [3] X. Zhao, H. Wang, M. Bai, Y. Xu, S. Dong, H. Rao, and W. Ming, “A comprehensive review of methods for hydrological forecasting based on deep learning,” *Water*, vol. 16, no. 10, p. 1407, 2024.
- [4] B. Kuhaneswaran, G. Sorwar, A. R. Alaei, and F. Tong, “Evolution of data-driven flood forecasting: Trends, technologies, and gaps—a systematic mapping study,” *Water*, vol. 17, no. 15, p. 2281, 2025.
- [5] M. S. B. M. Anik, C. An, and S. S. Li, “Evolution from the physical process-based approaches to machine learning approaches to predicting urban floods: a literature review,” *Environmental Systems Research*, vol. 14, no. 1, p. 15, 2025.
- [6] K. P. V. Robles, J. G. Solmerin, G. C. E. Pugat, and C. E. F. Monjardin, “A review of the advances and emerging approaches in hydrological forecasting: From traditional to ai-powered models,” *Water*, vol. 18, no. 1, p. 119, 2026.
- [7] A. Mosavi, P. Ozturk, and K.-w. Chau, “Flood prediction using machine learning models: Literature review,” *Water*, vol. 10, no. 11, p. 1536, 2018.
- [8] A. Msigwa and A. S. Makinde, “Application of machine learning techniques in hydrometeorological event prediction,” *Modeling and Monitoring Extreme Hydrometeorological Events*, pp. 135–161, 2024.
- [9] K. Xu, Z. Han, H. Xu, and L. Bin, “Rapid prediction model for urban floods based on a light gradient boosting machine approach and hydrological–hydraulic model: Xu et al.,” *International Journal of Disaster Risk Science*, vol. 14, no. 1, pp. 79–97, 2023.

- [10] Y. Wu, Z. Zhang, X. Qi, W. Hu, and S. Si, "Prediction of flood sensitivity based on logistic regression, extreme gradient boosting, and random forest modeling methods," *Water Science & Technology*, vol. 89, no. 10, pp. 2605–2624, 2024.
- [11] Y. Alwathaf, A. M. Al-Areeq, Y. A. Al-Masnay, A. R. Al-Aizari, and N. M. Al-Areeq, "Enhancing flood susceptibility mapping in sana'a, yemen with random forest and extreme gradient boosting algorithms," *Geocarto International*, vol. 40, no. 1, p. 2482707, 2025.
- [12] M. Jahanbani, M. H. Vahidnia, H. Aghamohammadi, and Z. Azizi, "Flood susceptibility mapping through geoinformatics and ensemble learning methods, with an emphasis on the adaboost-decision tree algorithm, in mazandaran, iran," *Earth Science Informatics*, vol. 17, no. 2, pp. 1433–1457, 2024.
- [13] M. Tosan, V. Nourani, O. Kisi, and M. Dastourani, "Evolution of ensemble machine learning approaches in water resources management: a review," *Earth Science Informatics*, vol. 18, no. 2, p. 416, 2025.
- [14] B. Du, M. Wang, J. Zhang, Y. Chen, and T. Wang, "Urban flood prediction based on pcswmm and stacking integrated learning model," *Natural Hazards*, vol. 121, no. 2, pp. 1971–1995, 2025.
- [15] N. M. Reddy and R. Srivastav, "Exploring the impact of optimization techniques on streamflow prediction using xgboost: a comparative analysis with satellite and reanalysis precipitation datasets," *Water Resources Management*, vol. 40, no. 1, p. 18, 2026.
- [16] X. Huang, Y. Geng, P. Liu, X. Hu, and Z. Wang, "Urban flood prediction using a hybrid xgboost-enhanced u-net model," *Journal of Hydrology*, p. 134822, 2025.
- [17] K. Namitha *et al.*, "Concept drift detection in data stream clustering and its application on weather data," *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, vol. 11, no. 1, pp. 67–85, 2020.
- [18] J. Yang, Z. Duan, Y. Chen, and J. Ou, "Assessing parametric rainfall models in reproducing tropical cyclone rainfall characteristics," *Atmospheric Research*, vol. 288, p. 106726, 2023.
- [19] D. S. Aquino, G. Gavier-Pizarro, A. J. Rescia, and R. D. Quintana, "Wetland responses to non-stationary hydro-climatic dynamics in the

- context of land cover and land use change,” *Remote Sensing Applications: Society and Environment*, vol. 34, p. 101156, 2024.
- [20] M. Tajgardoony, *Understanding and Quantifying Dataset Shift in Clinical Predictive Modeling*. PhD thesis, University of Pittsburgh, 2025.
 - [21] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, Z. Wang, S. Ebrahimi, and H. Wang, “Continual learning of large language models: A comprehensive survey,” *ACM Computing Surveys*, vol. 58, no. 5, pp. 1–42, 2025.
 - [22] G. Hovakimyan and J. M. Bravo, “Evolving strategies in machine learning: a systematic review of concept drift detection,” *Information*, vol. 15, no. 12, p. 786, 2024.
 - [23] S. Arora, R. Rani, and N. Saxena, “A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 4, p. e1536, 2024.
 - [24] M. Han, Z. Chen, M. Li, H. Wu, and X. Zhang, “A survey of active and passive concept drift handling methods,” *Computational Intelligence*, vol. 38, no. 4, pp. 1492–1535, 2022.
 - [25] B. R. Neupane and D. Pradhananga, “Observation to openness: Governing hydrological data in a non-stationary world,” *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 16, pp. 65–84, 2026.
 - [26] M. Gholizade, H. Soltanizadeh, M. Rahmanimanesh, and S. S. Sana, “A review of recent advances and strategies in transfer learning,” *International Journal of System Assurance Engineering and Management*, vol. 16, no. 3, pp. 1123–1162, 2025.
 - [27] Q. Gu and Q. Dai, “A novel active multi-source transfer learning algorithm for time series forecasting: A novel active multi-source transfer learning algorithm for time series forecasting,” *Applied Intelligence*, vol. 51, no. 3, pp. 1326–1350, 2021.
 - [28] L. C. Torres, L. M. Pereira, and M. H. Amini, “A survey on optimal transport for machine learning: Theory and applications,” *arXiv preprint arXiv:2106.01963*, 2021.
 - [29] R. Flamary, N. Courty, and M. Cuturi, “Optimal transport for machine learning,” *AG GDR ISIS*, 2017.

- [30] A. Khamis, R. Tsuchida, M. Tarek, V. Rolland, and L. Petersson, “Scalable optimal transport methods in machine learning: A contemporary survey,” *IEEE transactions on pattern analysis and machine intelligence*, 2024.
- [31] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty, “Optimal transport for structured data with application on graphs,” *arXiv preprint arXiv:1805.09114*, 2018.
- [32] G. Csurka, T. M. Hospedales, M. Salzmann, and T. Tommasi, *Visual domain adaptation in the deep learning era*. Morgan & Claypool Publishers, 2022.
- [33] N. Byaruhanga, D. Kibirige, S. Gokool, and G. Mkhonta, “Evolution of flood prediction and forecasting models for flood early warning systems: A scoping review,” *Water*, vol. 16, no. 13, p. 1763, 2024.
- [34] Q.-T. Tran, N.-A. Le-Khac, and M. Bertolotto, “Concept drift detection in image data stream: a survey on current literature, limitations and future directions,” *Artificial Intelligence Review*, vol. 59, no. 1, p. 33, 2025.
- [35] S. K. Mishra and P. Patri, “Flood forecasting using artificial intelligence (ai): A prisma-guided systematic review,” *Economic and Regional Studies/Studia Ekonomiczne i Regionalne*, vol. 19, no. 2, pp. 345–366, 2026.
- [36] R. K. Thomas, “Concept drift and non-stationarity in machine learning—an information-theoretic view,” *Journal of Tianjin University Science and Technology*, 2026.