Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable Flood Probability Modelling from Multidimensional Drivers Using Two-Stage Feature Selection and Hybrid Optimised Machine Learning

Saeid Najjar-Ghabel

saeid.najjarghabel@uwl.ac.uk

University of West London

Kourosh Behzadian

University of West London

Behnam Sadaghat

University of Tabriz

Research Article

Keywords: Flood Probability, Explainable machine learning, Recursive feature elimination, Hyperparameter optimisation, SHapley Additive exPlanations

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11454058/v1

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

Additional Declarations: No competing interests reported.

Interpretable Flood Probability Modelling from Multidimensional Drivers Using Two-Stage Feature Selection and Hybrid Optimised Machine Learning

Abstract

Floods are among the most destructive natural disasters, driven by environmental, infrastructural and socio-political factors. This paper presents an interpretable framework for estimating flood probability from 20 input variables in a benchmark dataset of 50,000 records, covering hydroclimatic and physical, land-use and ecosystem, infrastructure and water-management, and socioeconomic, governance and planning factors. To overcome the black-box nature of machine-learning models, the framework combines two-stage feature selection, in which Variance Inflation Factor (VIF) screening checks for collinearity and Recursive Feature Elimination (RFE) removes redundant predictors, with SHapley Additive exPlanations (SHAP)-based sensitivity analysis. Decision Tree Regression (DTR), Support Vector Regression (SVR) and Extreme Gradient Boosting (XGBoost) serve as base regressors, with Lungs Performance-Based Optimisation (LPO) used for hyperparameter tuning. VIF screening showed no severe multicollinearity (VIF < 6), so all 20 variables were retained, and RFE then retained the 18 highest-ranked predictors for each regressor. XGBoost was the best base regressor. Coupling each regressor with LPO improved test-set performance, increasing R^2 by 11-17% and reducing the root mean square error (RMSE) by 34-61%. The LPO-optimised XGBoost hybrid (XGLP) performed best (test $R^2 = 0.97$; RMSE = 0.0089), a 17% increase in R^2 and a 61% reduction in RMSE relative to the base XGBoost. SHAP analysis showed near-uniform contributions across predictors, consistent with the equal weighting of drivers in the benchmark index. The framework offers a transparent, reproducible pipeline applicable to observed flood records to support climate-resilience and infrastructure planning.

Keywords: Flood Probability; Explainable machine learning; Recursive feature elimination; Hyperparameter optimisation; SHapley Additive exPlanations.

1. Introduction

Floods remain one of the most devastating and widespread natural disasters, affecting millions of people annually and causing significant economic, social, and environmental damage (Fatica et al. 2024). Their increasing frequency and intensity are closely linked to a combination of natural processes and human activities (Nguyen et al. 2024). As climate change accelerates and urban landscapes expand, understanding and mitigating flood risks has become more complex than ever (Han et al. 2024). An effective response

requires not only predictive accuracy but also a deep understanding of the diverse and interconnected factors that shape flood dynamics (Rosenzweig et al. 2024).

At the heart of flood occurrences lies a web of interactions between environmental, infrastructural, and socio-political elements. Natural events such as monsoon systems, particularly their duration, intensity, and geographical shifts, can act as primary catalysts for flooding, especially in regions with seasonal rainfall patterns (Bazai et al. 2025). However, these meteorological triggers often interact with human-induced landscape transformations. Widespread deforestation, for instance, reduces the land's capacity to absorb water, increases surface runoff, and destabilises riverbanks (Potić et al. 2022). Likewise, rapid land-use changes (Andreadis et al. 2022) and unregulated urban sprawl (Alshammari et al. 2023) can obstruct natural water pathways, contributing to more frequent and severe flooding events.

Drainage infrastructure plays a critical mediating role in this context (Monachese et al. 2025). The design, capacity, and maintenance of drainage systems determine whether excess water is safely redirected or instead accumulates into destructive floodwaters (Wu et al. 2025). In many urban areas, outdated or poorly managed drainage networks amplify the risks posed by even moderate rainfall (Yang et al. 2023). Urban zoning policies further influence exposure and vulnerability; when residential or commercial development encroaches on flood-prone zones, often due to economic or political pressures, the manner in which land is allocated, regulated, and enforced reflects broader planning practices that directly affect flood resilience (Der Sarkissian et al. 2022).

Moreover, governance and institutional frameworks significantly shape both flood risk and flood response (Ibrahim et al. 2023). The quality of policy implementation, transparency in environmental regulation, disaster preparedness strategies, and the degree of inter-agency coordination are all essential components in managing flood hazards (Moises et al. 2024). In many regions, governance failures, such as inadequate risk assessments, corruption in construction planning, or a lack of public engagement, exacerbate flood vulnerability (Zhou et al. 2024). These socio-political dimensions are often less visible in technical analyses, yet they are central to understanding the full scope of flood probability (Carvalho and Spataru 2023). Capturing the complexity of such multi-dimensional systems calls for analytical approaches capable of processing and relating vast, heterogeneous information (Pradhan et al. 2023). Machine learning, with its capacity to identify patterns and interactions across diverse datasets, offers a promising path in this regard (Sharma et al. 2024). However, a major concern with many machine learning applications lies in their opacity: traditional models often deliver predictions without revealing the logic behind them (Wang et al. 2025). This has given rise to explainable machine learning techniques that not only predict outcomes but also expose the underlying factors contributing to them, bridging the gap between computational power and interpretive clarity (Hrast Essenfelder et al. 2025). Such approaches enable a more transparent and

informed understanding of complex phenomena like flood risk, fostering better planning and decision-making across disciplines.

Machine-learning methods provide a flexible means of representing nonlinear relationships and complex interactions among heterogeneous flood-related variables (Pramanik et al. 2026). Methods such as decision tree, support vector machine, and gradient-boosting are among the most popular ones used for mapping flood susceptibility and estimating flood index. The effectiveness of these methods, however, is not only limited using several irrelevant predictors, suboptimal configurations of hyperparameters, overfitting and information leakage during the construction of a model but may also be compromised by the quality of data. Moreover, several highly performing machine-learning models lack the ability to clearly describe how they make their predictions (Li et al. 2025). Explainable AI provides tools through which one can trace the contribution of each input variable to the output of a machine-learning model. In particular, SHapley Additive exPlanations (SHAP) are a method of measuring the influence of each variable on a prediction of a model, either positively or negatively, with reference to a starting value. This type of explanation helps detect errors in the model, promotes understandability and reveals non-linear relations (Hrast Essenfelder et al. 2025).

Recent studies demonstrate the growing application of machine learning to different flood-assessment problems. Islam and Chowdhury (2024), for example, compared random forest and support vector machine models for flash-flood susceptibility assessment in northeastern Bangladesh. Sseguya and Jun (2024) evaluated feedforward, convolutional, and long short-term memory networks within a deep neural-network ensemble. Other investigations have applied logistic regression, decision trees, k-nearest neighbours, random forests, support vector machines, XGBoost, Gaussian-process regression, and metaheuristic hybrid models to flood susceptibility and flood-potential assessment in different geographical settings (Vojtek et al. 2023; Taromideh et al. 2024; Demissie et al. 2024).

Table 1 summarises the most recently developed models (2022-2024) for flood risk assessment, together with the data, main results and reference of each study. A closer examination of the predictors, feature-selection, optimisation and explainability steps used in these studies reveals three recurring limitations, each of which is addressed in the present study. First, as the key predictors reported in Table 1 indicate, the predictors used are almost exclusively physical (topography, hydrology, rainfall, soil and land cover), whereas infrastructural drivers (e.g., drainage capacity, dam quality, urbanisation) and socio-political drivers (e.g., governance quality, encroachment, disaster preparedness) are absent. In contrast, the present study considers 20 predictors covering hydroclimatic, land-use, infrastructural and socio-political governance domains. Second, feature refinement and hyperparameter optimisation are rarely combined: only Singha et al. (2024) applied formal feature selection (Boruta), and only Vojtek et al. (2023) used a

metaheuristic optimiser (a genetic algorithm). None combined multicollinearity screening, model-specific recursive feature elimination and metaheuristic tuning in a single pipeline, which is the approach adopted in the present study. Third, apart from Singha et al. (2024), none of the reviewed studies used SHAP to quantify the direction and magnitude of each predictor's contribution. Moreover, the recently proposed lungs performance-based optimisation (LPO) has not yet been applied to flood modelling. To the best of our knowledge, no recent study has integrated physical, infrastructural and governance drivers with VIF-RFE refinement, LPO-based tuning and SHAP attribution to improve both the accuracy and the interpretability of flood-probability estimation.

Table 1. Recently developed ML-driven models for flood risk assessment

Method	Data	Results	Reference
RF and SVM (statistical + spatial metrics)	Local flash-flood records in NE Bangladesh; 5 zones	RF outperformed SVM: AUC = 0.964, acc = 92.1%, κ = 0.894; overall RF–SVM map agreement = 73.3% (very-low class: RF-SVM = 97.15%, SVM-RF = 81.54%)	Islam and Chowdhury (2024)
FFNN, CNN, LSTM + DNNE	Nile basin (1993-2022; 6 predictors)	DNNE best; LSTM: 99.82 % (AUC = 0.99), FFNN: 99.63 % (AUC = 0.99), CNN: 95.42 % (AUC = 0.89)	Sseguya and Jun (2024)
LR, RF, SVM, KNN, AdaBoost and XGBoost on 13 GIS predictors	1,528 flood + 528 non-flood points (1996-present)	XGBoost & RF best: acc = 0.82-0.97 across resolutions; key predictors: stream distance, soil type, rainfall, elevation, impervious surfaces; stacking improved robustness	Demissie et al. (2024)
DT, KNN, LR, SVM, RF (tuned)	Semarang: 2,884 points; 14 factors	KNN delivered highest zoning accuracy; top factors: curve number, distance to river, building density	Yuwono et al. (2024)
Ensemble vs. CART and RF	Urban flood predictors	Ensemble outperformed CART & RF in accuracy and bias reduction; generated credible urban flood-hazard maps	Taromideh et al. (2024)
GPR and SVM w/ 4 kernels	Koshi basin 2010-2020; 13 factors	SVM-PUK best ($r^2 \approx 0.995$; RMSE ≈ 0.051); classified 31.00 % high & 1.10 % very high flood-risk area	Baig et al. (2022)
10 ML models + Boruta/SHAP	Arambag geospatial data	RF AUC = 0.847; AdaBoost = 0.839; SHAP/Boruta flagged precipitation, elevation, river distance; 17.2-18.6 % area highly susceptible	Singha et al. (2024)
FFNN, XGBoost and GA-hybrids	Gidra basin: 10 k points; 11 factors	GA-XGBoost best (AUC = 0.97 vs. XGBoost = 0.96); hybrids improved high-risk detection to 12.3 % vs. 11.0 % in stand-alone models	Vojtek et al. (2023)

Abbreviations: AdaBoost: adaptive boosting; AUC: area under the receiver operating characteristic curve; acc: accuracy; CART: classification and regression tree; CNN: convolutional neural network; DNNE: deep neural network ensemble; DT: decision tree; FFNN: feedforward neural network; GA: genetic algorithm; GPR: Gaussian process regression; KNN: k-nearest neighbours; LR: logistic regression; LSTM: long short-term memory; ML: machine learning; NE: north-eastern; PUK: Pearson VII universal kernel; RF: random forest; SHAP, SHapley Additive exPlanations; SVM: support vector machine; XGBoost: extreme gradient boosting; κ: Cohen's kappa coefficient.

This study aims to develop and evaluate an accurate and explainable hybrid machine-learning framework for estimating a composite flood-probability index from multidimensional environmental, infrastructural,

land-use and institutional variables. The specific objectives are to: (1) determine whether VIF-RFE feature refinement improves model parsimony and predictive performance; (2) compare baseline and LPO-optimised DTR, SVR, and XGBoost models; (3) identify the predictors that contribute most strongly to the estimated index; and (4) examine the extent to which the modelling framework can reproduce the composite response represented in the benchmark dataset.

2. Methodology

This study developed an explainability-oriented machine-learning framework for estimating a constructed composite flood-probability index from multiple environmental, infrastructural, demographic, and governance-related factors. The framework comprised four principal stages as shown in Fig. 1, including a multidimensional flood-driver dataset; feature refinement using VIF and recursive feature elimination (RFE); flood-probability modelling using decision tree regression (DTR), support vector regression (SVR), and extreme gradient boosting (XGBoost); and hyperparameter optimisation using lungs performance-based optimisation (LPO). The resulting hybrid models were subsequently evaluated through five-fold cross-validation and interpreted using SHapley Additive exPlanations (SHAP).

We initially trained the three conventional regressors using their baseline configurations to establish reference performance. We then coupled LPO separately with each regressor to generate the LPO-DTR, LPO-SVR, and LPO-XGBoost hybrid models. We assessed model performance using error-, bias-, dispersion, and agreement-based indicators. Finally, SHAP analysis was applied to the best-performing model to quantify the direction and magnitude of each predictor's contribution to the estimated flood probability.

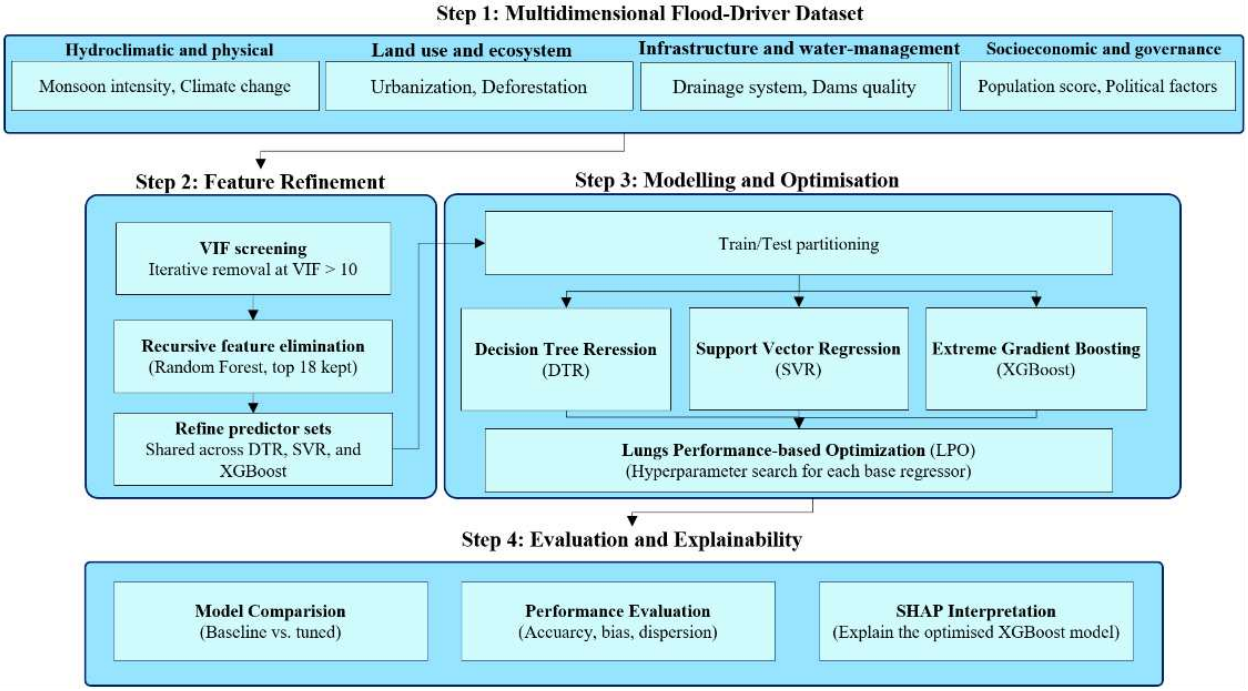


Fig. 1 The framework for LPO-based flood-probability modelling

2.1. Data Collection and Predictor Variables

The data used in this study were obtained from the publicly available Flood Prediction Dataset (Khalid 2024). The dataset consisted of 50,000 instances, characterised by 20 predictor variables and one target variable: Flood Probability (FP). The 20 predictors and their abbreviations are grouped in Table 2.

Table 2. Predictor groups and variables

Predictor group	Variables and abbreviations
Hydroclimatic and physical factors	Monsoon Intensity (MI); Climate Change (CC); Topography Drainage (TD); Siltation (Si); Coastal Vulnerability (CV); Landslides (Ls); Watersheds (Ws)
Land use and ecosystem factors	Wetland Loss (WL); Deforestation (DF); Urbanization (Ur); Encroachments (En); Agricultural Practices (AP)
Infrastructure and water-management factors	Drainage Systems (DS); Dam Quality (DQ); River Management (RM); Deteriorating Infrastructure (DI)
Socioeconomic, Governance and planning factors	Population Score (PS); Political Factors (PF); Ineffective Disaster Preparedness (IDP); Inadequate Planning (IP)

Prior to model development, the dataset was thoroughly randomised. This random permutation eliminates potential bias from the original data ordering and ensures that the training, validation, and testing splits are representative. Randomisation is a standard procedure to prevent unintended patterns from influencing

model training and evaluation. Fig. 2 displays the dendrogram resulting from feature clustering. To elucidate variable relationships, we applied agglomerative hierarchical clustering to the 20 predictors. Variables from similar domains typically form clusters. For instance, *UR* and *WL* merge at a low linkage distance, indicating a shared association with land development. The results, as visualised in Fig. 2, indicate distinct predictor groups: urban and developmental factors on one side, climate and hydrological factors on the other, and intermediate clusters that bridge these domains.

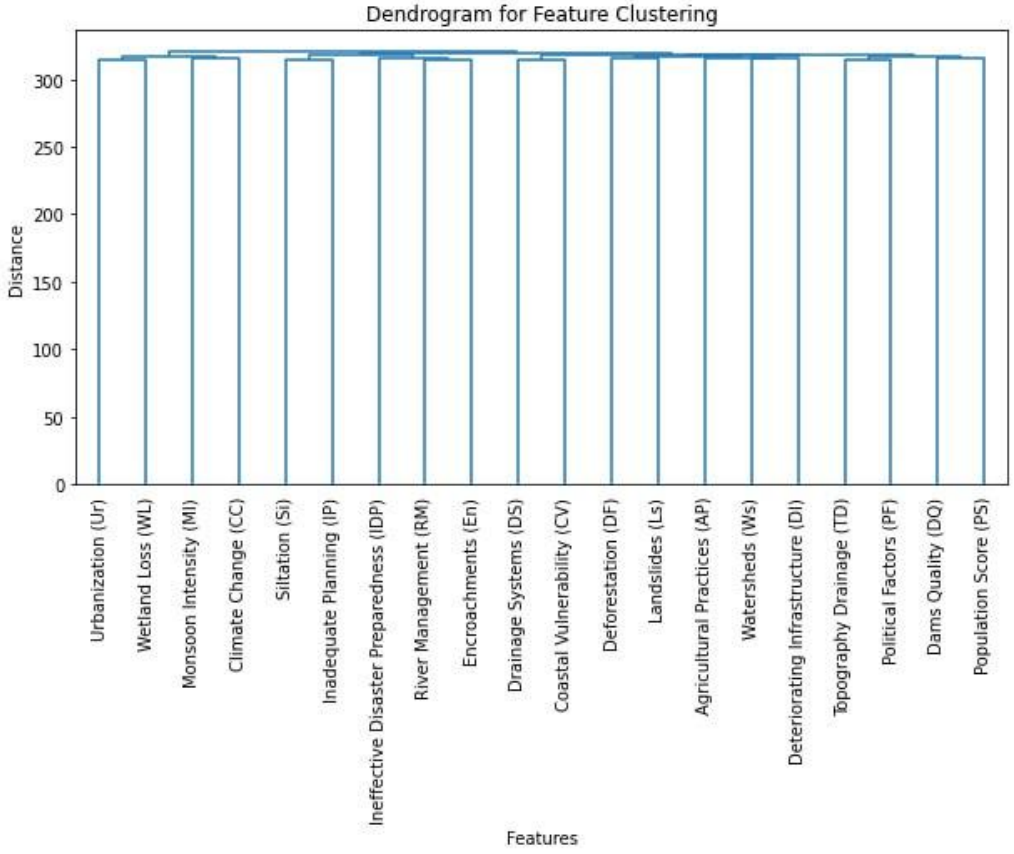


Fig. 2 Dendrogram showing feature clustering

2.2. Predictor Assessment and Feature Refinement

In the data collection phase, multicollinearity was assessed by calculating the variance inflation factor (VIF) for each independent variable. Multicollinearity is identified when two or more predictors exhibit a high degree of correlation, a condition that can artificially inflate or deflate the standard errors of estimated coefficients and consequently render truly significant effects nonsignificant (Alin 2010). Multicollinearity was evaluated using VIF before model training. For each predictor X_j, the variance inflation factor is calculated as Equation (1):

$$VIF_j = \frac{1}{1 - R_j^2} \quad (1)$$

Where VIF_j is the variance inflation factor for predictor X_j and R_j^2 is obtained by regressing X_j on all the other independent variables. An intercept was included in each auxiliary regression to avoid uncentred VIF estimates. In addition, we centred all predictors before calculating VIFs.

A VIF value of 1 corresponds to a regressor that is independent of the other regressors; a VIF value increasing beyond 1 implies increased multicollinearity. VIF over 5 should indicate potential collinearity problems worth considering more carefully, whereas a VIF over 10 should show clear evidence of serious collinearity problems. Predictors, however, were not eliminated only because they were higher than some arbitrary number, but instead their conceptual significance, redundancy, and effect on model stability were also evaluated alongside the other measures, which is in keeping with the view that VIF cutoffs must be read against the background (O'brien 2007).

After the multicollinearity check, the Recursive Feature Elimination (RFE) method was applied separately to DTR, SVR, and XGBoost models. RFE is a one-step-at-a-time procedure in which the feature's importance is evaluated based on the selected importance metric, and the least important one is removed. At the very beginning, the model is built based on all the features at hand, and then for each feature an importance index (e.g. the absolute value of regression coefficients in linear models or built-in importance scores of tree models) is computed. The least important predictor is dropped, and the remaining features are assessed again following the same steps. This process of fitting, scoring, and removing continues till the number of features reaches the required number of dimensions. Thus, the main purpose of this procedure is to eliminate all features except those that are extremely important for the model's predictions (Brownlee 2020). For each regressor, the 18 highest-ranked predictors were retained for subsequent model development and hyperparameter tuning.

2.3. Machine Learning Regression Models

2.3.1. Decision Tree Regression (DTR)

The Decision Tree Regression (DTR) algorithm is a type of supervised learning that is popular and is equally well suited for regression and classification tasks. A decision tree algorithm creates a set of hierarchical decision nodes and leaf nodes that at each level divides the dataset based on some specific criteria. When it is used for regression, the DTR makes use of a set of input features to model a response, predicting values that span a continuous range, instead of using categories which are well defined (Erdal 2013; Ahmad et al. 2021). The biggest feature of the DTR model is that it can divide the space of input variables into a number of partitions that are mutually exclusive and collectively exhaustive. For each such split, the model creates a corresponding leaf to which one of the records is assigned through the decision

path, and the result corresponds with the label of the leaf node to be retrieved. Through a process of splitting, stopping, and pruning, the DTR is essentially trying to build a model that will accurately predict and will be able to be reasonably understood by a human, that reveals the hidden patterns in a given dataset through a suitable partitioning of the data.

2.3.2. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning technique aimed at forecasting continuous outputs from past data. It works by identifying a function that most closely fits the data within a given tolerance, the so-called error margin, which allows slight deviations from the true values without imposing penalties. An extended method based on SVR, called nu-SVR, aims to adjust epsilon and the level of prediction accuracy according to features of your data. The nu-SVR tries to reduce prediction errors at a level that also controls the complexity of the model so that the result is an accurate and generalisable forecast. In most cases, the forecasting technique based on SVR comprises several steps: getting data from different reliable sources, preparing the data, for example by normalisation or transformation, in order to make the data suitable for training a learner, and splitting the whole data set into a set for training the model and a set for testing. Training involves the model getting to know the data and extracting the parameters that best describe it; then testing evaluates the model's correctness in different aspects. Once the model is verified, it is used to forecast new values. SVR is therefore well suited to problems that involve nonlinearities or noise (Li et al. 2009).

2.3.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a tree-boosting framework that can scale well. This framework unites algorithmic improvements with system-level design to deliver top-tier performance and efficiency. Besides, it adopts regularised learning so that one aspect in the learning process, model accuracy, is not at the expense of another aspect, model complexity, through utilising not only first-order gradients but second-order gradients as well. The use of gradients is helpful in making very good and quick split decisions. When it comes to very big data, the weighted quantile sketch introduced by XGBoost provides a good level of accuracy in split proposals, whereas the sparsity-aware split-finding algorithm handles the missing/ zeros values efficiently by not requiring the data to be dense, thereby resulting in significant gains in speed, especially in the cases where onehot-encoded features are involved. At the system level, XGBoost makes use of features such as cache-aware column blocks, block compression and sharding in order to support learning directly from the hard disk while also enabling parallel and distributed execution, which together allow the processing of billions of instances on anything from a single laptop to large computer clusters (Chen and Guestrin 2016).

2.4. Lungs Performance-Based Optimisation (LPO)

The predictive performance of DTR, SVR, and XGBoost depends strongly on their hyperparameter configurations. Lung's performance-based optimisation (LPO) was therefore employed separately for each regression model to identify a suitable configuration. LPO is a population-based metaheuristic inspired by respiratory impedance, which comprises respiratory resistance and respiratory reactance. In the algorithm, these concepts are translated into complementary search mechanisms intended to balance exploration of the hyperparameter space with refinement around promising solutions (Ghasemi et al. 2024)

For model m, the i-th candidate configuration at iteration t is represented by a complete vector of model-specific hyperparameters:

$$\theta_{m,i}^{(t)} = \left[\theta_{m,i,1}^{(t)}, \theta_{m,i,2}^{(t)}, \ldots, \theta_{m,i,D_m}^{(t)}\right]^T \quad (2)$$

Here, $\theta_{m,i}^{(t)}$ represents the hyperparameter candidate vector i for model m at iteration t, D_m is the number of hyperparameters optimized for model m. The initial value of hyperparameter j for candidate i was sampled within its predefined search boundaries:

$$\theta_{m,i,j}^{(0)} = L_{m,j} + u_{i,j} \cdot \left(U_{m,j} - L_{m,j}\right) \quad (3)$$

In this expression, $\theta_{m,i,j}^{(0)}$ is the value of the j-th hyperparameter inside vector $\theta_{m,i}^{(t)}$, $L_{m,j}$ and $U_{m,j}$ are the lower and upper bounds of hyperparameter j, respectively, and $u_{i,j} \sim U(0,1)$ is a random variable drawn

from a standard uniform distribution. At each iteration, the LPO search operators generated a trial vector, denoted by $u_{m,i}^{(t)}$. Each trial element was restricted to its admissible interval using:

$$\theta_{m,i,j}^{\text{trial}} = \min\left(U_{m,j}, \max\left(L_{m,j}, \theta_{m,i,j}^{\text{trial}}\right)\right) \quad (4)$$

Integer-valued hyperparameters were rounded to the nearest integer, whereas continuous hyperparameters retained their numerical values. Greedy selection was subsequently applied so that a trial configuration replaced the current candidate only when it reduced the objective function:

$$\theta_{m,i}^{(t+1)} = \begin{cases} u_{m,i}^{(t)}, & \text{if } f\left(u_{m,i}^{(t)}\right) < f\left(\theta_{m,i}^{(t)}\right) \\ \theta_{m,i}^{(t)}, & \text{otherwise} \end{cases} \quad (5)$$

LPO was applied independently to DTR, SVR, and XGBoost. For XGBoost, the search space included the number of estimators and maximum tree depth. For DTR, maximum tree depth was optimised, while for SVR, the penalty parameter C was optimized.

The fitness of every candidate configuration was evaluated using three-fold inner cross-validation. For model m, candidate configuration $\theta_{m,i}$, and validation fold k, the root mean square error (RMSE) was calculated as Equation 6:

$$\text{RMSE}_{m,i,k} = \sqrt{\frac{1}{N_k}\sum_{p=1}^{N_k}\left(y_{k,p} - \hat{y}_{m,i,k,p}\right)^2} \quad (6)$$

Here, N_k is the number of observations in the validation fold k, $y_{k,p}$ is the observed composite flood-probability index, and $\hat{y}_{m,i,k,p}$ is the corresponding value predicted by the model m. The overall LPO objective function $f\left(\theta_{m,i}\right)$ is defined as the mean validation RMSE across all folds shown in Equation (7):

$$f\left(\theta_{m,i}\right) = \frac{1}{3}\sum_{k=1}^{3} RMSE_{m,i,k} \quad (7)$$

Consequently, the selected optimal hyperparameter configuration θ_m^* for model m was determined by Equation 8:

$$\theta_m^* = \arg\min_{\theta_m \in \Omega_m} f(\theta_m) \quad (8)$$

Where Ω_m denotes the admissible hyperparameter search space. Therefore, a lower objective-function value indicated a better candidate configuration. After optimisation, each model was refitted using the complete development dataset, its selected predictor subset, and the optimal hyperparameter configuration θ_m^*. Final performance was subsequently evaluated using an independent test dataset.

2.5. Evaluation Metrics

Model performance was assessed using the complementary metrics summarised in Table 3. Together, these metrics evaluate prediction accuracy, systematic bias, residual variability, and agreement between measured and predicted values (Contributors 2024).

Table 3. Summary of the statistical metrics used to evaluate model performance.

Metrics	Equation	Description
Root Mean Square Error (*RMSE*)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2}$	Measures the overall magnitude of prediction errors while assigning greater weight to large deviations
Mean Absolute Error (*MAE*)	$\frac{1}{n}\sum_{i=1}^{n}\lvert X_i - Y_i\rvert$	-
Percentage Deviation ($\boldsymbol{PD_i}$)	$\frac{X_i - Y_i}{X_i} * 100$	Expresses the signed percentage error for the *i* th observation
Average Percentage Deviation (*APD*)	$\frac{1}{n}\sum_{i=1}^{n} PD_i$	Measures the mean signed percentage error and indicates systematic model bias
Average Absolute Percentage Deviation (*AAPD*)	$\frac{1}{n}\sum_{i=1}^{n}\lvert PD_i\rvert$	-
Standard Deviation of Residuals (*SD*)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$	Quantifies the variability of prediction errors around their mean
Pearson Correlation Coefficient ($\boldsymbol{R}$)	$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})}}$	Measures the strength and direction of the linear relationship between measured and predicted values
Coefficient of Determination ($\boldsymbol{R^2}$)	$1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	Indicates the proportion of variation in the measured values explained by the model

Notes: n is the number of observations; X_i and Y_i are the measured and predicted values, respectively; $D_i = X_i - Y_i$ is the residual; and $\bar{X}$, $\bar{Y}$, and $\bar{D}$ are the corresponding mean values.

2.6. Five-Fold Cross-Validation

Fig. 3 depicts a five-fold cross-validation (CV) approach in which the original dataset is first randomised to avoid potential order effects and split randomly into five equal-sized subsets (folds). In each of the five CV loops, a different fold is used as the test case, whereas the other four serve as the training set. After training on the four folds, the model is tested on the remaining fold, so each observation ends up once as a test case and four times as model input. After the five loops, the performance metrics from the folds are collated, usually by calculating the mean and standard deviation to provide the most accurate assessment of the model's ability to generalise. Not only does this method reduce the risk of variance caused by a single splitting of train and test sets, but it also serves as a measure of overfitting resistance by testing the model multiple times with different and varied data partitions (Kohavi 1995).

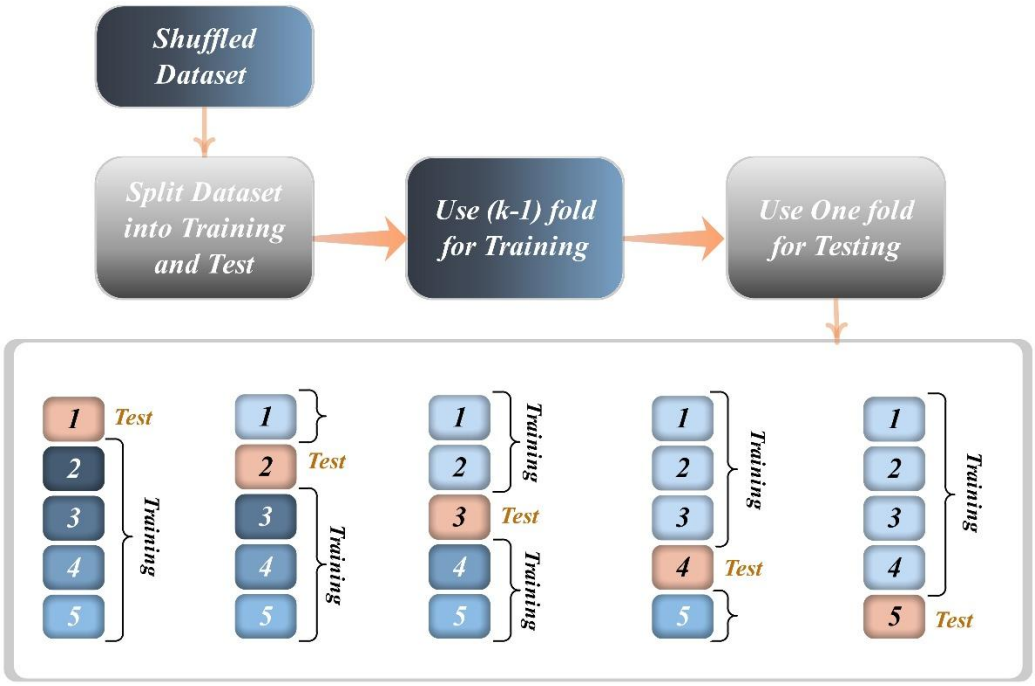


Fig. 3 Visual illustration of the dataset split according to the 5-fold cross-validation

2.7. Feature Importance

The Shapley Additive explanations (SHAP) methodology is used to determine the contribution of individual features to a particular prediction. Thus, the model is not only explained, but feature importance can also be determined. SHAP values are calculated based on concepts from cooperative game theory. In this way, feature contributions are fairly allocated across all possible feature combinations, resulting in consistent and highly transparent feature relevance scores. Contrary to traditional significance indicators, such as gain measures in tree models, which only describe the model behaviour at the single data level, SHAP provides information at two levels: the local explanation of the prediction and the global overview of feature effects (Granados et al. 2023).

3. Results and Discussion

3.1. Feature Assessment and Model Validation

3.1.1. Multicollinearity Check and Feature Selection

The VIF values in Fig. 4 are very close (about 5.65 to 5.76), with the highest (~5.75) and the lowest (~5.66). As expected, none of the VIFs exceeds 10, so multicollinearity is not very high. One way to assess whether the data are well-conditioned is by using variance inflation factors (VIFs). A variance inflation factor of 10 generally indicates a rather severe level of multicollinearity. VIF values over 10 suggest close examination of multicollinearity, and in extreme cases, removal of variables may be necessary. Therefore, if multicollinearity is moderate, it may be enough to be aware of it without removing variables.

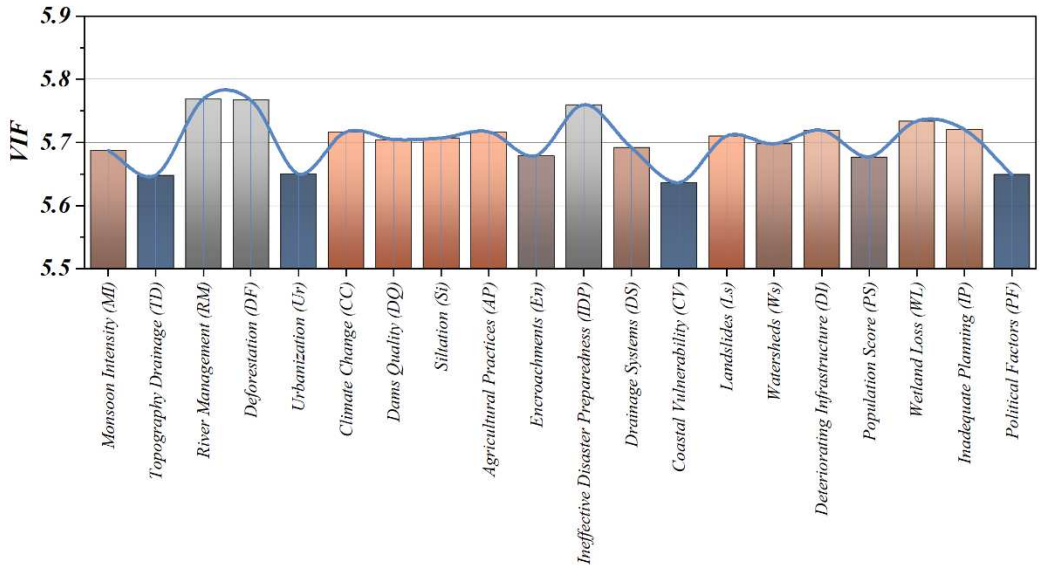


Fig. 4 Multicollinearity by VIF score

RFE-based feature selection was conducted with three regressors: DTR, XGBoost, and SVR. Table 4 presents the final features chosen with each of them, while the full elimination order is detailed in Tables A1-1 through A1-3 in the Appendix 1. Each model produced a different ranking of predictor importance. With DTR, Monsoon Intensity (MI) is kept as the top one and even survives the last iteration. With XGBoost, Population Score (PS) becomes the main predictor. For SVR, on the other hand, River Management (RM) is one of the most important variables. The models also differ in the selection of secondary features; e.g., DTR's 2nd- and 3rd-ranked variables are Deteriorating Infrastructure (DI) and Inadequate Planning (IP), whereas the remaining predictors are ordered differently.

Notably, MI, PS, and RM were each the most important feature in one model, but no single "core" feature was the most important in all models. The recursive elimination process showed some common points: certain features (like AP and Si) were removed from all model sequences initially, suggesting they were not important. Meanwhile, MI in DTR, PS in XGBoost, and RM in SVR were the last to be removed, outlasting most other features in their respective models. These findings (shown in Table 4 and Table A1, Table A2, and Table A3 in Appendix 1) suggest that, with different feature-selection methods, each modelling method emphasises a different set of predictors, and this does not reflect a set of commonly important features.

Table 4. Feature selection of the models based on the RFE ranking method.

ID	Variable	RFE Rank		
		DTR	XGBoost	SVR
1	Monsoon Intensity (MI)	1	20	18
2	Topography Drainage (TD)	10	19	9

3	River Management (RM)	12	18	1
4	Deforestation (DF)	13	17	8
5	Urbanisation (Ur)	17	16	11
6	Climate Change (CC)	7	15	13
7	Dams Quality (DQ)	8	14	6
8	Siltation (Si)	6	13	17
9	Agricultural Practices (AP)	15	12	2
10	Encroachments (En)	14	11	5
11	Ineffective Disaster Preparedness (IDP)	11	10	12
12	Drainage Systems (DS)	16	8	15
13	Coastal Vulnerability (CV)	19	9	7
14	Landslides (Ls)	5	7	3
15	Watersheds (Ws)	18	6	16
16	Deteriorating Infrastructure (DI)	2	5	20
17	Population Score (PS)	20	1	14
18	Wetland Loss (WL)	9	2	19
19	Inadequate Planning (IP)	3	3	10
20	Political Factors (PF)	4	4	4

3.1.2. Correlation of Variables and Cross-Validation

We further investigated the linear correlation between each predictor and the target variable and summarised it in Tables A4-A6 of Appendix 2. Overall, FP was weakly correlated with individual variables, since all the Pearson correlation coefficients were roughly less than 0.23, indicating that no single factor was an overwhelming linear predictor. For example, from the results of DTR-based analysis displayed in Table A5, the highest correlation coefficient for FP is approximately 0.229. In the XGBoost-based correlation table (Table A4), the highest correlation is slightly less, from about 0.18 to 0.20 for the top-ranked features. SVR correlations (Table A6), like DTR correlations and those based on XGBoost, are generally not very high, but all show very weak and nearly zero FP correlations. For example, the correlations never exceed 0.18. Although FP is to some degree associated with a few predictors such as River Management (RM) or Deteriorating Infrastructure (DI), which sit at the upper end of this band, the associations are generally weak and vary by algorithm, highlighting the importance of considering multiple factors jointly rather than relying on any single variable.

Fig. 5 presents the results of a test that each model achieved on each of the five setups with five-fold cross-validation. For instance, the XGBoost model attained its score of 0.839 on Fold 2. Whereas both SVR and DTR models reached their top performances during Fold 5, which resulted in scores of 0.827 and 0.829,

respectively. In addition, they are the maximum values recorded for predictive accuracy of these three models through five-fold cross-validation. All the model's performance scores on the remaining cross-validation folds are below these two results. These outcomes demonstrate that the XGBoost was, on the whole, better at identifying patterns and producing more accurate predictions when the data were available from the earlier splits. On the contrary, SVR and DTR are better when the splits are later. Overall, XGBoost produced the best prediction in Fold 2 (the initial partition), while the SVR and DTR models had the best outcomes in the final data split (Fold 5).

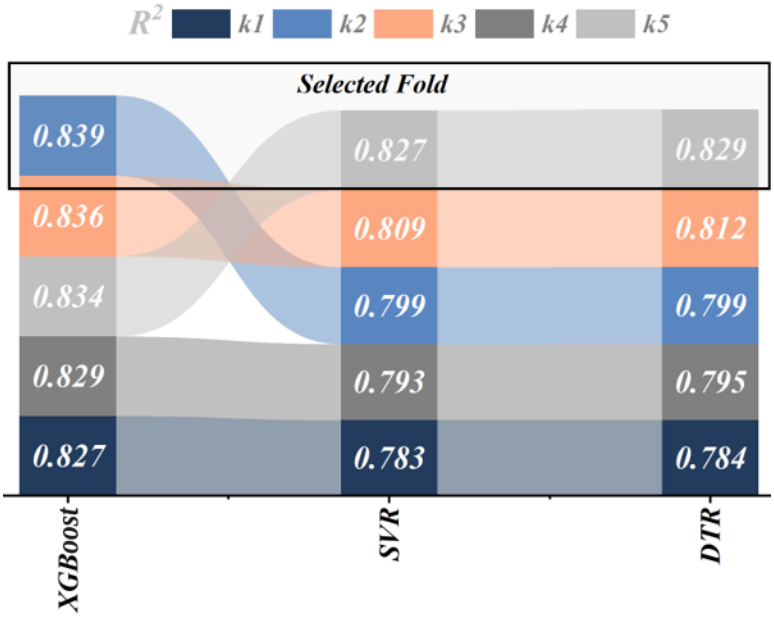


Fig. 5 Results of 5-fold cross-validation

3.2. Hyperparameter Tuning and Convergence Curves

Combining their respective machine learning model with the LPO algorithm to automatically select optimal settings for hyperparameters is the basis for all those newly developed models: XGLP (XGBoost + LPO), SVLP (SVR + LPO), and DTLP (DTR + LPO). The algorithm works so effectively that it is possible to maintain both aspects of exploration and exploitation without bias in the tuning. The settings for each model's hyperparameters, described in detail in Table 5, differ significantly in their strategy and nature. At a high level, XGLP needed a large ensemble (69 trees) and deep trees (max_depth= 58) to discover nonlinear relationships. In contrast, SVLP needed a low level of regularisation (C = 57). This was because they were aware of the risks of model overfitting and wanted to strike a fair balance between precision and generalisation. DTLP, on the other hand, was a more straightforward approach with a much smaller individual tree size (max_depth = 27).

The convergence curves in Fig. 6 reveal the RMSE changes with every iteration of 200 steps, thus demonstrating how LPO effectively leads every hybrid model to strong optima. The XGLP model decreased its RMSE most sharply at the beginning and was able not only to maintain this but even to reach as low as 0.0087 at the end. This shows very clearly its capability of capturing a highly complicated mapping relationship with a very small number of residuals. SVLP took a while to reach a final RMSE value of 0.0156. This reflects how difficult it is to optimise kernel-based regressors in the situation when the number of dimensions is very high. DTLP, having the RMSE of 0.0129 midway between the other two, showed a smooth, steady rate of change in RMSE, which can mean that DTLP is the model of choice when moderate problem complexity is at stake. This difference in performance at the end seems to be very closely related to setting and selecting hyperparameters: XGLP's decision trees, being deep and expansive, have naturally contributed to a higher level of accuracy; DTLP, however, due to its regularisation and its relatively simpler architecture, introduced certain limits on model flexibility. These results, taken together, make it very convincing that customised hyperparameter tuning is necessary for maximum model performance and that a hybrid loss penalisation with different models (here LPO) is a flexible approach that may be applied across different algorithmic frameworks.

Table 5. Hyperparameter properties of the developed hybrid models.

Models Name	Hyperparameter	Description	Range	Optimum Value
XGLP	n_estimators	number of trees in the ensemble	[8, 78]	69
	max_depth	maximum levels per tree	[4, 78]	58
SVLP	C	inverse strength of regularisation	[7, 79]	57
DTLP	max_depth	maximum levels per tree	[3, 79]	27

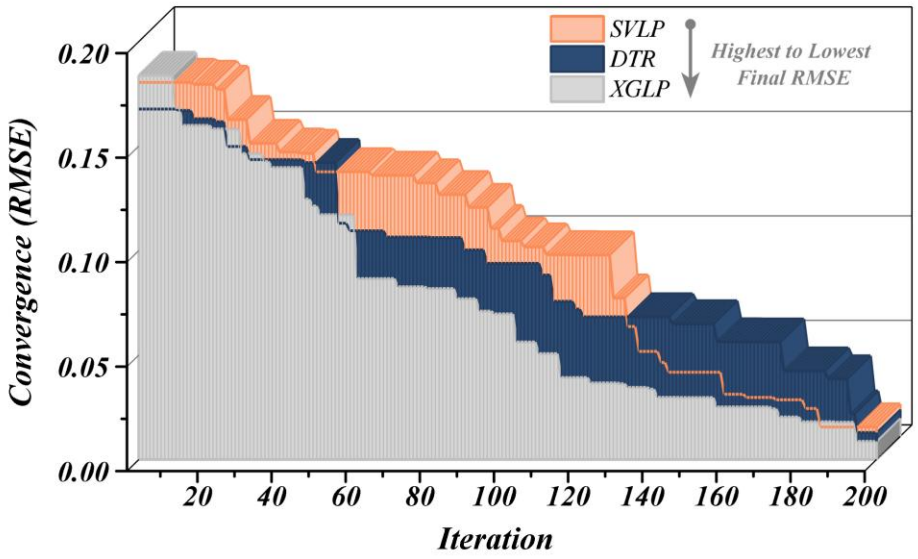


Fig. 6 Convergence curves of the developed hybrid models across 200 iterations

3.3. Performance Evaluation of the Models

The performance results of the base models across various training, validation, and testing phases of the task are given in Table 6 in multiple metrics. XGBoost is shown to have higher predictive power and accuracy and was the top performer in all the phases, with train: 0.8327, validation: 0.8332 and test: 0.8278 as its highest values, respectively. It clearly outperformed SVR by about 1.6% in the training session and 1.1% in the testing session, and it was able to outperform DTR by a little margin in training performance by 0.6% and test performance by 0.2%.

It is important to note that validation of DTR was a bit better than XGBoost (R^2 = 0.8337 vs. 0.8332, respectively), and that their generalisation was quite similar. However, DTR still slightly performed worse than XGBoost, with the test performance of 0.8263 versus XGBoost's 0.8278 being a 0.2% gap. Moreover, SVR remained far behind, as the test accuracy was 0.8187 for them, which is approximately 1.1% less than XGBoost and 0.9% below DTR. XGBoost was the most effective in terms of error metrics. RMSE (test: 0.0227) and MAE (test: 0.0183) of XGBoost were the lowest out of the three. Moreover, when comparing with SVR, the reduction in RMSE was as high as 3.4% and 4.4% in terms of MAE, while for DTR, it improved by 2.2% and 0.5%, respectively.

SVR's performance was significantly lower, with RMSE at 0.0235 and MAE at 0.0191, implying that the model lacks precision, unlike XGBoost, which has a much higher precision. DTR, however, has intermediate errors indicating moderate precision, which aligns with the simplicity and balanced structural

model of DTR. The bias in model predictions was reflected in the APD figures. For XGBoost, there was a very slight tendency for overestimation in training (0.0020) and testing (0.0621). On the other hand, SVR generally underestimated during the whole phase from training (-0.0200) to testing (-0.0505). DTR had a mixed trend, with an underestimation of -0.0833 in testing only. With the lowest AAPD, XGBoost (test: 3.7008), DTR followed as the second one (3.6685), and SVR ended up last one (3.8562). Compared to SVR, XGBoost reduced the AAPD by 4%.

Like RMSE, the standard deviation (SD) of errors reflected a comparable trend, with XGBoost exhibiting the smallest (0.0227) SD and closely followed by DT Regression (DTR) with 0.0228 SD, while at the same time both were much better than SVR, whose SD was 0.0235. Further validation using correlation coefficients (R), XGBoost outperformed SVR and DTR with a high value of 0.9099 test correlation. It was followed closely by SVR and DTR, with the gap between XGBoost being about 0.6% over SVR and only 0.1% over DTR.

Based on the above discussion and comparison results, overall, XGBoost stands out to be the most accurate and stable model, whereas, the SVR with a lesser degree of accuracy because of regularisation has emerged as the weakest one and the Decision Tree Regression as an alternative model offers the compromise between easy-to-understand and interpretability features, on one hand, and performance, on the other hand, by nearly optimal results. It may be said that such findings illustrate the necessity to keep a balance between error minimisation, model complexity and generalisation when it comes to predicting flood probabilities.

Table 6. Performance results of the base models across the train, validation, and test phases.

Criteria	XGBoost			SVR			DTR		
	Train	Val.	Test	Train	Val.	Test	Train	Val.	Test
R^2	0.8327	0.8332	0.8278	0.8197	0.8177	0.8187	0.8276	0.8337	0.8263
RMSE	0.0224	0.0223	0.0227	0.0234	0.0234	0.0235	0.0228	0.0223	0.0228
MAE	0.0180	0.0180	0.0183	0.0189	0.0189	0.0191	0.0183	0.0179	0.0182
APD	0.0020	0.0407	0.0621	-0.0200	0.0808	-0.0505	0.0090	0.0800	-0.0833
AAPD	3.6466	3.6494	3.7008	3.8240	3.8230	3.8562	3.7064	3.6238	3.6685
SD	0.0224	0.0223	0.0227	0.0234	0.0234	0.0235	0.0228	0.0223	0.0228
R	0.9125	0.9128	0.9099	0.9054	0.9043	0.9048	0.9097	0.9131	0.9090

Performance of the hybrid models at the three different stages (training, validation and testing) is shown in terms of predictive accuracy in Table 7. As it is evident from the table, among the three hybrid models (XGLP, SVLP and DTLP), the XGLP model proved to be most accurate on all three testing phases. With

prediction accuracies of 0.9704 (training), 0.9714 (validation), and 0.9692 (testing), the XGLP hybrid model has surpassed the SVLP by around 6.0% on the testing phase and the DTLP by around 3.4%. SVLP performed the poorest by having the lowest prediction accuracy on all three phases (training: 0.9115; validation: 0.9112; testing: 0.9108). DTLP, on the other hand, showed medium performance in prediction accuracy (training: 0.9374; validation: 0.9372; testing: 0.9363).

From an error point, one can further observe the difference between the models. XGLP achieved the best performances in term of Root Mean Square Error (RMSE) (training: 0.0087; validation: 0.0086; testing: 0.0089) and Mean Absolute Error (MAE) (training: 0.0069; validation: 0.0068; testing: 0.0071), being DTLP a 31% lower in test RMSE (0.013 vs. 0.0089) and SVLP a 44% higher (0.0156 vs. 0.0089). Interestingly, DTLP (errors: test RMSE = 0.0130; test MAE = 0.0104) errors were considerably lower than those for SVLP (errors: test RMSE = 0.0156; test MAE = 0.0126), indicating that the model has an efficient way to deal with complexity and trade-off error.

Error analysis in terms of percentage also reveals the high reliability and stable accuracy of XGLP on the testing set (bias in terms of APD was 0.0119, and AAPD was 1.4349%). For SVLP, on the other hand, the average prediction underestimated (APD is -0.0374), and it shows an even higher AAPD (2.5521%). DTLP, in addition, shows a mixed bias (APD is -0.0495); it has AAPD (2.1003%), which is lower. Interestingly, the deviations in error mirrored RMSE trends, such that XGLP's SD (test: 0.0089) is much stricter than that of DTLP (0.0130) and SVLP's (0.0156).

Finally, XGLP's strong correlation with test results (R = 0.9845) exceeding DTLP by 1.7% and SVLP by 3.0%. The overall results suggest that the XGLP hybrid model provides better accuracy, very little bias, and a reliable error distribution, whilst DTLP is a good balance between accuracy and interpretability, and SVLP, despite being better than its base SVR model, remains the least precise of the hybrids.

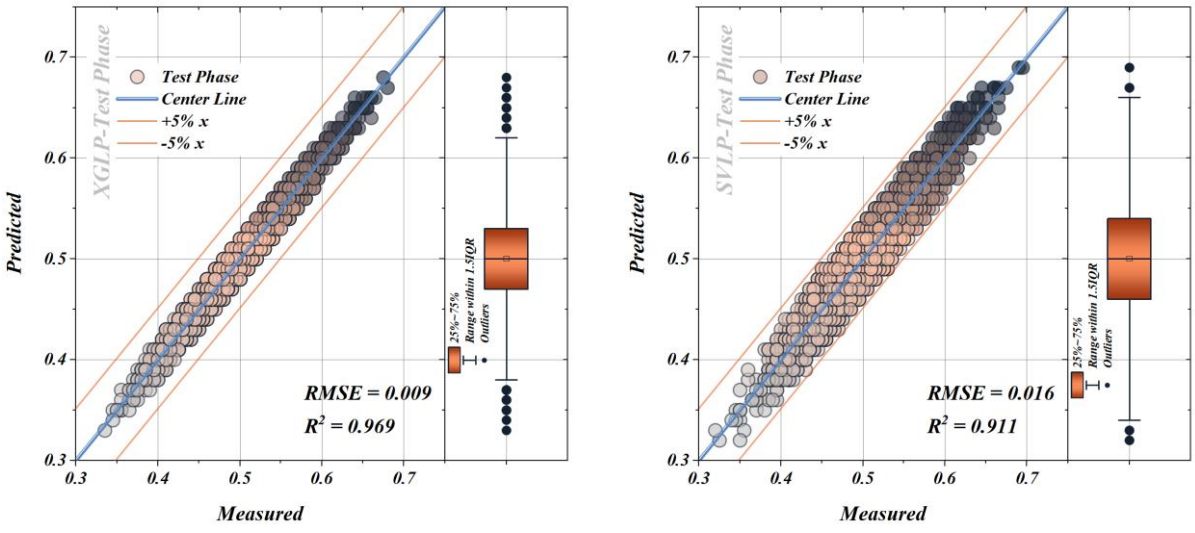
Table 7. Performance results of the hybrid models across the train, validation, and test phases.

Criteria	XGLP			SVLP			DTLP		
	Train	Val.	Test	Train	Val.	Test	Train	Val.	Test
R^2	0.9704	0.9714	0.9692	0.9115	0.9112	0.9108	0.9374	0.9372	0.9363
RMSE	0.0087	0.0086	0.0089	0.0156	0.0156	0.0156	0.0129	0.0129	0.0130
MAE	0.0069	0.0068	0.0071	0.0126	0.0126	0.0126	0.0104	0.0104	0.0104
APD	-0.0004	-0.0165	0.0119	-0.0019	0.0535	-0.0374	-0.0021	0.0487	-0.0495
AAPD	1.4034	1.3834	1.4349	2.5517	2.5538	2.5521	2.0962	2.1035	2.1003
SD	0.0087	0.0086	0.0089	0.0156	0.0156	0.0156	0.0129	0.0129	0.0130
R	0.9851	0.9856	0.9845	0.9547	0.9546	0.9544	0.9682	0.9681	0.9676

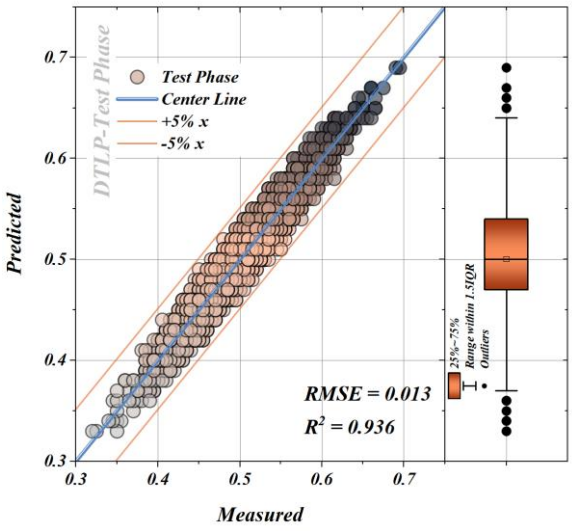
Fig. 7 shows the scatter plots and box-and-whisker summaries of the predictions vs. measurements of the three hybrid models during the test phase. In XGLP's chart, the scatter points almost exactly follow the 1:1 line without deviation, and all the points lie within the ±5 % error bands. Its boxplot shows an interquartile range (IQR) that is very narrow, and the median is almost aligned with the bottom of the box. Only a very few observations fall outside the whiskers as outliers, which confirms that XGLP is a very unbiased and consistent model.

In SVLP's plot, the scatter points are more spread out around the 1:1 line so that the ±5 % boundaries are slightly crossed by a number of observations. This increased degree of variability is reflected in the boxplot, whose IQR is twice as large as the one corresponding to XGLP, while its median is only a tiny bit higher up the line, i.e., it indicates that the tendency towards positive bias may be a very minor one indeed. Besides, there are also some points that are very clearly outliers, which show that the SVLP has made predictions outside of the ±5 % tolerance limits on a number of occasions.

On DTLP's chart, the scatter points cover a range that is wider than one that contains ±5 % of the total points, but most of them fit under the ±5% corridor, and only a small proportion of them scatter widely. The boxplots' IQR of this group is wider than that of XGLP, yet narrower than that of SVLP. Its median is almost right on the line, but the whiskers go to the ends of the error limits. Besides, there are also a good number of points just outside the whiskers, which show rare but very short deviations from perfect predictions.



(a) (b)



(c)

Fig. 7 Scatter plots of the hybrid models in the testing phase, along with box plots. (a) XGLP, (b) SVLP, and (c) DTLP

Fig. 8 depicts relative prediction errors for the hybrid models in box, scatter, whisker, and violin plots. At the very least, the plot is informative: 1) XGLP has the smallest interquartile range per cent of error, as demonstrated by the thin blue box at its centre. Its median line is almost right in the middle of the error axis, which means that the overall prediction is nearly unbiased, if not completely unbiased. Besides this fact, the violin shape is very tightly drawn around the median, indicating a strong concentration. It could be said that the 5% error whiskers cover most of the data, with only a few mild data points going beyond, confirming that XGLP has both precision and stability of the highest class. 2)SVLP, on the other hand, has an error distribution which is quite a lot wider. The interquartile range of the model is roughly twice that of XGLP, and from a glance, the spread of errors in both positive and negative directions is very evident. The median is just a little over the zero line; this indicates a small positive prediction bias. Moreover, SVLP has one or two very large outliers which go beyond a ±9% error line, thus this hybrid has a good potential compared to pure SVR but still produces predictions which are not as stable as the other hybrids. 3)DTLP is somewhere in the middle. The IQR from the grey boxplot of this plot is wider than that of XGLP's but narrower than that of SVLP, besides the latter the median is almost perfectly located at the zero point; this means no bias or very small bias. The error whisker length is close to ±7%, i.e., there are a few bigger deviations from XGLP's predictions in this set, but only one or two very large deviations are from SVLP. Finally, the violin shape error distribution is confirmed as moderate, i.e., the majority are ±5% error, but as we get closer to each end of the distribution, we notice more variability.

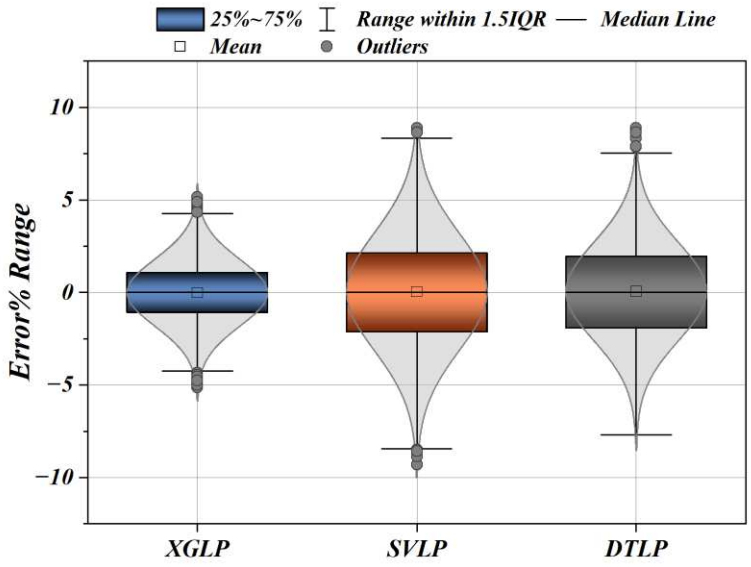


Fig. 8 Error box plots of the hybrid models

Fig. 9 overlays the REC curves of each base model with its corresponding hybrid variant and reports the area under each curve (AUC) as a concise summary of error-tolerance performance. By plotting the fraction of predictions whose absolute deviation falls below a given threshold, these curves reveal how rapidly each model accumulates "acceptable" predictions as the tolerance widens: a steeper ascent and a leftward shift indicate a higher concentration of low-error forecasts.

In the REC curve of the XGBoost versus XGLP, the hybrid XGLP trace rises more sharply, reaching ~99 % of predictions within a 0.05 deviation, whereas the base XGBoost roughly reaches ~0.85% of accuracy within 0.05 deviation. Both curves converge to full accuracy, but in different deviations; XGLP's consistent lead across the error spectrum yields an AUC of 0.990, with its sharp increase in accuracy. This clear leftward displacement demonstrates that LPO tuning densifies the most precise forecasts, pushing a larger share of XGLP's predictions into the smallest-error band.

In the REC curve of SVR versus SVLP, the SVLP trace climbs more steeply, achieving approximately 90 % of predictions within a 0.04 deviation, whereas the base SVR only reaches about ~78 % accuracy at that same threshold. Both curves eventually approach 98 % coverage by roughly a 0.9 deviation. This leftward shift confirms that LPO optimisation concentrates SVLP's errors more tightly around zero, yielding far fewer large deviations than the unoptimized SVR. Similarly, in the REC comparison of DTR and DTLP, the hybrid DTLP curve accelerates more quickly: it attains close to 98 % of predictions within a 0.04 deviation, while DTR manages only about 85 % at the same point. The clear gap between DTLP and DTR throughout the error range highlights how LPO tuning sharpens the model's precision, shifting a larger fraction of predictions into narrow-error bands. Across all three model families, the REC analysis confirms

that the hybrids consistently shift the error-tolerance profiles leftward and steepen the initial rise of their REC curves. The higher AUC values of XGLP (0.990), SVLP (0.986), and DTLP (0.987) compared with their base models demonstrate that hyperparameter optimisation via LPO not only improves average accuracy metrics but also markedly increases the proportion of highly precise predictions under stringent deviation criteria.

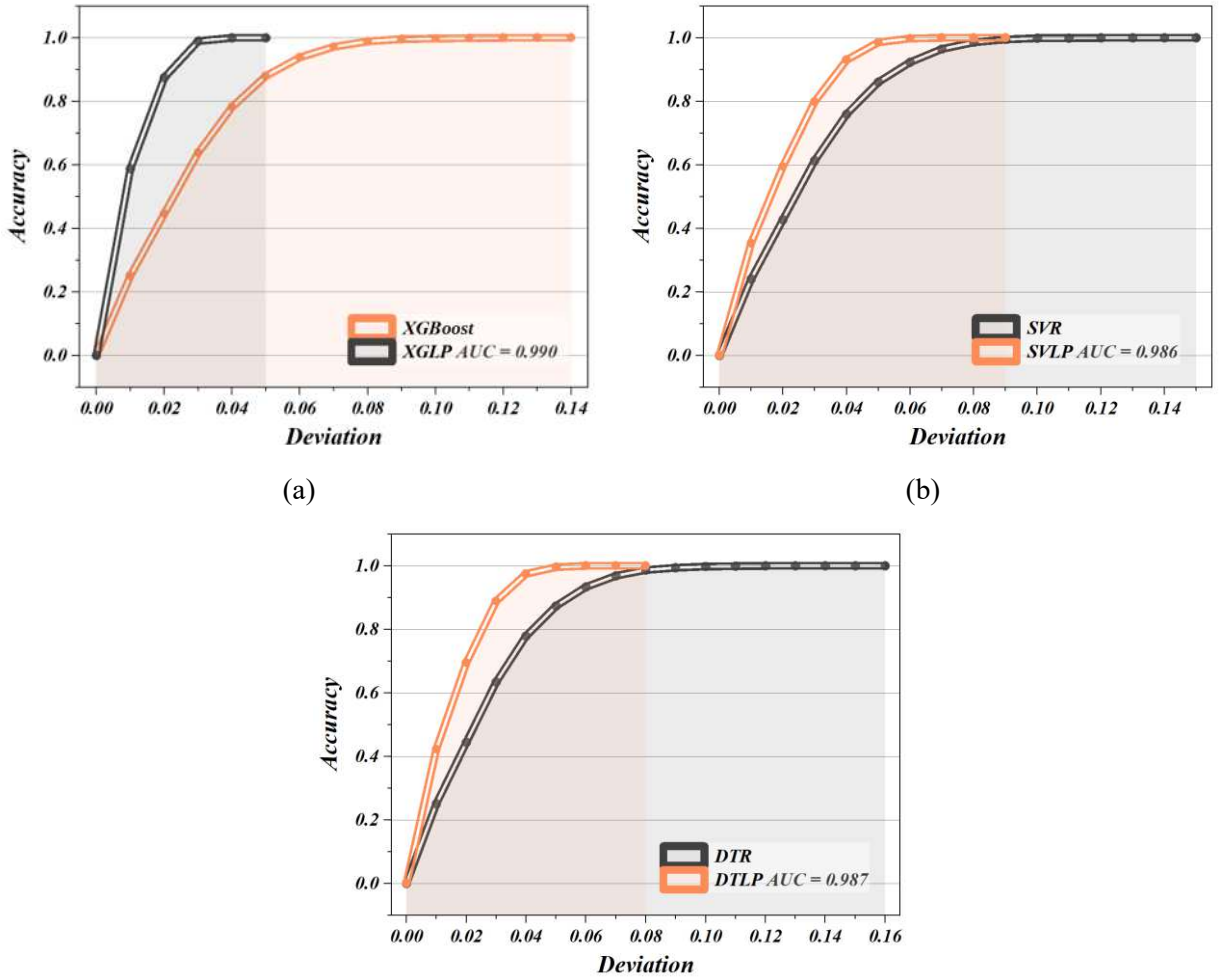


Fig. 9 The REC curve of the models, along with AUC values

Table 8 highlights the differences in the true flood probability distribution and the distribution of the hybrid models' predictions at four value ranges, reporting for each the average, minimum, and maximum values. For XGLP, the overall average prediction (0.5001) lies just 0.0001 (0.02 %) below the measured average (0.5002). The minimum predicted value of 0.3300 underestimates the observed minimum of 0.3350 by 0.0050 (1.49 %), while the maximum matches exactly at 0.6800. Within the 0.3-0.4 band, XGLP's mean of 0.3789 is 0.0027 (0.71 %) lower than the measured 0.3816; in the 0.4-0.5 band its mean of 0.4605 trails

0.4625 by 0.0020 (0.43 %). In the 0.5-0.6 range, XGLP's 0.5318 is 0.0020 (0.37 %) below the observed 0.5338, and for values above 0.6, its 0.6159 falls short of 0.6183 by 0.0024 (0.39 %).

SVLP's average prediction (0.5003) is 0.0002 (0.04 %) above the measured average of 0.5001. Its minimum of 0.3200 exactly equals the observed 0.3200, while its maximum of 0.6900 underestimates the true maximum of 0.6950 by 0.0050 (0.72 %). In the 0.3-0.4 interval, SVLP's mean of 0.3764 is 0.0037 (0.97 %) below the measured 0.3801; in the 0.4-0.5 band its 0.4591 is 0.0029 (0.63 %) lower than the observed 0.4620. Within 0.5-0.6, SVLP's 0.5332 is almost identical to 0.5339 (a 0.0007 or 0.13 % difference), and above 0.6, its 0.6207 slightly exceeds 0.6189 by 0.0018 (0.29 %).

DTLP yields an average of 0.5004, 0.0003 (0.06 %) above the measured 0.5001. Its minimum of 0.3300 overestimates the true 0.3200 by 0.0100 (3.13 %), and its maximum of 0.6900 underestimates the observed 0.6950 by 0.0050 (0.72 %). In the lowest band, DTLP's 0.3762 is 0.0039 (1.03 %) below the measured 0.3801; in the 0.4-0.5 range, its 0.4592 trails 0.4620 by 0.0028 (0.61 %). For 0.5-0.6, DTLP's 0.5331 is 0.0008 (0.15 %) under the observed 0.5339, and above 0.6, its 0.6195 is 0.0006 (0.10 %) above the measured 0.6189.

Table 8. Statistical properties of the base and hybrid models.

Model name	Statistical Analysis						
	Avg	Min	Max	Mean in range			
				0.3 to 0.4	0.4 to 0.5	0.5 to 0.6	More than 0.6
Measured	0.5002	0.3350	0.6800	0.3816	0.4625	0.5338	0.6183
XGLP	0.5001	0.3300	0.6800	0.3789	0.4605	0.5318	0.6159
Measured	0.5001	0.3200	0.6950	0.3801	0.4620	0.5339	0.6189
SVLP	0.5003	0.3200	0.6900	0.3764	0.4591	0.5332	0.6207
Measured	0.5001	0.3200	0.6950	0.3801	0.4620	0.5339	0.6189
DTLP	0.5004	0.3300	0.6900	0.3762	0.4592	0.5331	0.6195

3.4. Sensitivity Analysis Results

The SHAP analysis in Fig. A1 (see Appendix 3) shows a consistent pattern across the predictor set. For every predictor, the SHAP values are centred close to zero and spread over a similar range (approximately -0.03 to +0.05), with near-identical, roughly symmetric distributions across the samples. No single predictor or group of predictors stands out with a markedly wider distribution, which agrees with the narrow range of the normalised mean absolute SHAP values (4.68-5.32%). This near-uniform contribution reflects the

structure of the benchmark target, a composite index that weights its constituent drivers almost equally, rather than a ranking of real-world flood drivers. The SHAP results therefore confirm that the fitted model reproduces this structure, and small differences in rank between predictors should not be over-interpreted.

3.5. Discussion

3.5.1. Comparative Model Performance and Interpretability

Beyond the hybridisation of various regressor models, the use of the LPO technique has further contributed immensely to the increase in the level of predicted power and confidence. Among all the models tested, XGLP proved to have the highest values for the test set's performance metrics; R^2 was 0.9692, while RMSE was merely 0.0089. Such a decrease in the value of RMSE would result in a reduction of the error rate from the baseline model, XGBoost, by only 60.8%. The fact that the standard error of XGLP is much less compared to that of XGBoost (0.0089 vs 0.0227), the difference in the variance of the model's predictiveness, also points out that XGLP's errors are much more concentrated around zero compared to those of XGBoost. The regression error characteristic (REC) curves further support and illustrate these results by indicating that the LPO-optimised models outperform the others in prediction accuracy, especially at tighter error margins. LPO has thus been found to be one of the most reliable ways one can identify a set of hyperparameters that go well with nonlinear and interdependent predictor features.

To interpretably maintain the pipeline, VIF screening, RFE, and SHAP have been integrated. The model XGLP is the hybrid that has greater structural complexity compared with DTLP, so by its nature it is less transparent. But SHAP compensates by providing insights both at the global (dataset-level) and local (instance-level) levels. Based on the sensitivity analysis reported in Section 3.4, the SHAP results do not reveal a particular group of drivers: the normalised mean importance ranges from 4.68 to 5.32% for the predictors. This is what would be expected, in fact, because of the nearly equal weighting of inputs in the composite flood-probability index. What SHAP does apart from ranking variables is providing directions. A variable may be very important, but its level of importance cannot tell you whether an increase in it would make prediction probability go up or down, so the signed SHAP values show it clearly. DTLP corresponds to another trade-off: retaining very good prediction ($R^2 = 0.9363$) within a single-tree structure, the paths of which can be traced from end to end. Given that the selected depth (max_depth = 27) would yield a tree too big so that the whole thing cannot be easily grasped, the transparent nature of a DTLP lies mainly in checking individual paths rather than the whole model. In other words, it might be a better option if auditability and quick deployment are more important than getting the maximum prediction accuracy.

4. Conclusions

The present paper has described the development of the integrated interpretable machine-learning framework for the estimation of the flood probability composite index. This framework has been built based on a benchmark dataset of 50,000 records and 20 variables which represent five flood-related domains: meteorological, hydrological, agricultural, infrastructural, and socioeconomic data. The methodology of the presented framework comprises four phases: the detection of possible colinear predictors and feature selection (by using the Variance Inflation Factor method), the use of stepwise removal approach as a recursive method of the feature selection, finding optimal values for ML model parameters (hyperparameter optimisation) by means of the Lungs Performance-based Optimisation (LPO) algorithm metaheuristic, and finally a method for the post-hoc attribution via SHAP. The following models have been tested both in their standard forms as well as in their optimised hybrid forms via LPO: Decision Tree Regression (DTR), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost). As a comparison, they have also been named DTLP, SVLP, and XGLP (the hybrids). On the held-out test set, amongst the three hybrids, XGLP had the highest scores ($R^2 = 0.9692$, RMSE = 0.0089, MAE = 0.0071, Pearson's R = 0.9845, AAPD = 1.4349%). In terms of the RMSE, XGLP's RMSE was improved by 44% and 31% upon a better model; we follow the same convention throughout this paper and all other percentage comparisons. DTLP's accuracy level was not so high ($R^2 = 0.9363$), although it was a single-tree algorithm in which each decision path could be traced, hence being more suitable for applications where the requirement of auditability supersedes that of gaining from marginal improvements in the predictive performance.

These results require an important qualification. In this benchmark, the target variable is a composite index built from the flood risk scores for the 20 predictors instead of an observed flood event. Basically, the index is just an averaged weighted sum of the 20 standardised predictors. An ordinary least squares regression model fitted on the same data set and tested on the same held-out test set would have easily uncovered the model behind it. Additionally, the SHAP analysis discussed in Section 3.4 is in line with this: the average of the normalised mean contribution lies within 4.68 and 5.32% for all 20 predictors, an almost uniform distribution, which can be explained by a model which gives equal weight to its underlying index instead of being driven by a differential influence among the drivers. This study demonstrates that a pipeline combining VIF screening, recursive feature elimination, metaheuristic hyperparameter tuning, and SHAP attribution can recover a known generative structure and expose both the magnitude and the direction of each predictor's contribution, which is a prerequisite for applying such a pipeline to data whose underlying structure is genuinely unknown. However, it is not proof that a real-world flood risk factor dominates the others or that such attributions should be taken as causal statements about flooding, but rather, they should be understood as properties of a fitted model and its training data.

The first of these limitations is that all training and evaluation have been conducted with the help of a single benchmark dataset whose target has been artificially constructed; hence, the reported accuracy figures only reveal performance of the pipeline under those particular synthetic conditions, and they are not to be interpreted as indicators of flood levels which have actually been recorded elsewhere or at other times. Another issue is that the dataset does not contain any spatial or event-level characteristics, which makes the framework unfit for being validated against inundation maps or operational forecasts. Furthermore, the predictors have not changed over time, and the models have only been tested on paper hydrometric measurements without real-time testing. In addition, a comparison between baseline and LPO-tuned models does not pinpoint the effect of LPO only, as the baseline models were left untuned; thus, the improvement shown may rather result from general tuning rather than from the optimiser.

Future research in this area should consider an event-based flood inventory with observed inundation outcomes for a real case study, to which the developed pipeline can be applied so that SHAP attributions can be checked against recorded floods. Future studies should also benchmark LPO against Bayesian optimisation, genetic algorithm and particle swarm optimisation, with all baselines tuned under the same computational budget, to isolate the contribution of the optimiser itself.

Appendix

Appendix 1

Table A1. Feature selection details for DTR based on RFE

0.049	***0.094***	***0.121***	***0.169***	***0.214***	***0.266***	***0.304***	***0.352***	***0.399***	***0.443***	***0.493***	***0.542***	***0.587***	***0.631***	***0.686***	***0.729***	***0.784***	***0.829***	***0.826***	***0.825***
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
		19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
			20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
				14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
					8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
						6	6	6	6	6	6	6	6	6	6	6	6	6	6
							7	7	7	7	7	7	7	7	7	7	7	7	7
								17	17	17	17	17	17	17	17	17	17	17	17
									2	2	2	2	2	2	2	2	2	2	2
										11	11	11	11	11	11	11	11	11	11

											3	3	3	3	3	3	3	3	3
												4	4	4	4	4	4	4	4
													10	10	10	10	10	10	10
														9	9	9	9	9	9
															12	12	12	12	12
																5	5	5	5
																	15	15	15
																		13	13
																			18

606

607 **Table A2**. Feature selection details for XGBoost based on RFE

0.050	***0.093***	***0.132***	***0.174***	***0.211***	***0.254***	***0.302***	***0.351***	***0.402***	***0.458***	***0.495***	***0.549***	***0.594***	***0.641***	***0.689***	***0.732***	***0.775***	***0.831***	***0.834***	***0.835***
17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
		19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
			20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
				16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
					15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
						14	14	14	14	14	14	14	14	14	14	14	14	14	14
							12	12	12	12	12	12	12	12	12	12	12	12	12
								13	13	13	13	13	13	13	13	13	13	13	13
									11	11	11	11	11	11	11	11	11	11	11
										10	10	10	10	10	10	10	10	10	10
											9	9	9	9	9	9	9	9	9
												8	8	8	8	8	8	8	8
													7	7	7	7	7	7	7
														6	6	6	6	6	6
															5	5	5	5	5
																4	4	4	4
																	3	3	3
																		2	2

608

609 **Table A3**. Feature selection details for SVR based on RFE

0.052	*0.094*	*0.127*	*0.176*	*0.213*	*0.261*	*0.317*	*0.352*	*0.396*	*0.453*	*0.498*	*0.548*	*0.593*	*0.642*	*0.688*	*0.741*	*0.780*	*0.827*	*0.830*	*0.826*
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
		14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
			20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
				10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
					7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
						13	13	13	13	13	13	13	13	13	13	13	13	13	13
							4	4	4	4	4	4	4	4	4	4	4	4	4
								2	2	2	2	2	2	2	2	2	2	2	2
									19	19	19	19	19	19	19	19	19	19	19
										5	5	5	5	5	5	5	5	5	5
											11	11	11	11	11	11	11	11	11
												6	6	6	6	6	6	6	6
													17	17	17	17	17	17	17
														12	12	12	12	12	12
															15	15	15	15	15
																8	8	8	8
																	1	1	1
																		18	18
																			16

610

611 **Appendix 2**

612 **Table A4**. Correlation of the variables within dataset concerning XGBoost model

Variable	*PS*	*WL*	*IP*	*PF*	*DI*	*Ws*	*Ls*	*DS*	*CV*	*IDP*	*En*	*AP*	*Si*	*DQ*	*CC*	*Ur*	*DF*	*RM*	*FP*

DQ	*Si*	*AP*	*En*	*IDP*	*CV*	*DS*	*Ls*	*Ws*	*DI*	*PF*	*IP*	*WL*	*PS*
0.007	-0.002	0.002	0.004	-0.004	0.002	0.000	0.002	0.001	0.001	0.007	-0.002	0.004	1
-0.001	0.008	-0.008	0.006	-0.004	-0.007	-0.006	-0.002	-0.008	0.000	0.003	-0.004	1	
0.002	0.012	-0.002	0.001	0.004	-0.010	-0.008	-0.009	0.009	0.007	-0.004	1		
0.002	-0.002	-0.001	-0.008	0.001	-0.003	0.001	0.004	-0.003	0.004	1			
-0.004	0.003	0.003	0.003	0.000	-0.005	-0.003	0.003	0.005	1				
0.005	0.004	0.004	0.004	0.000	0.002	-0.003	0.004	1					
0.000	-0.002	0.004	-0.003	-0.002	0.003	0.004	1						
0.004	-0.008	0.001	-0.006	0.004	0.008	1							
-0.010	-0.004	-0.004	-0.007	0.001	1								
-0.001	0.001	0.003	0.001	1									
0.002	0.000	-0.007	1										
0.002	-0.004	1											
0.003	1												
1													

613

614

Table A5. Correlation of the variables within dataset concerning DTR model

Variable	*FP*	*RM*	*DF*	*Ur*	*CC*
MI	0.227	-0.001	0.000	0.002	-0.008
DI	0.224	0.002	-0.001	0.013	-0.001
IP	0.223	0.000	-0.001	-0.001	0.000
PF	0.225	-0.006	0.001	-0.002	-0.004
Ls	0.229	0.001	-0.002	0.004	0.003
Si	0.228	0.000	0.002	-0.003	0.003
CC	0.223	-0.004	0.005	-0.001	0.001
DQ	0.218	0.002	-0.001	-0.005	-0.008
PS	0.215	0.000	-0.003	-0.001	-0.001
TD	0.225	0.006	0.002	0.001	-0.005
IDP	0.218	0.009	-0.004	-0.011	0.000
RM	0.222	0.004	0.003	-0.001	-0.003
DF	0.227	0.000	-0.001	-0.001	0.002
En	0.227	0.008	-0.001	0.000	-0.003
AP	0.221	0.007	0.001	0.007	1
DS	0.221	-0.010	-0.011	1	
Ur	0.220	0.005	1		
Ws	0.229	1			
FP	1				

Ls	*PF*	*IP*	*DI*	*MI*
-0.002	0.005	0.003	0.002	1
0.003	0.004	0.007	1	
-0.009	-0.004	1		
0.004	1			
1				

DS	*AP*	*En*	*DF*	*RM*	*IDP*	*TD*	*PS*	*DQ*	*CC*	*Si*
0.000	0.003	-0.004	-0.005	0.002	0.002	-0.003	-0.004	0.003	0.006	-0.003
-0.003	0.003	0.003	-0.002	0.001	0.000	0.004	0.001	-0.004	0.003	0.003
-0.008	-0.002	0.001	-0.001	0.000	0.004	0.005	-0.002	0.002	0.000	0.012
0.001	-0.001	-0.008	0.001	-0.006	0.001	0.008	0.007	0.002	-0.004	-0.002
0.004	0.004	-0.003	0.005	-0.004	-0.002	-0.002	0.002	0.000	0.001	-0.002
-0.008	-0.004	0.000	-0.001	0.000	0.001	0.011	-0.002	0.003	0.002	1
-0.008	-0.003	0.000	0.001	0.007	-0.005	-0.002	-0.008	-0.003	1	
0.004	0.002	0.002	-0.001	0.008	-0.001	-0.004	0.007	1		
0.000	0.002	0.004	0.000	-0.001	-0.004	0.006	1			
-0.002	-0.005	-0.007	0.002	0.002	0.001	1				
0.004	0.003	0.001	0.002	0.006	1					
0.002	0.004	0.009	0.005	1						
-0.001	0.003	-0.004	1							
-0.006	-0.007	1								
0.001	1									
1										

615

616

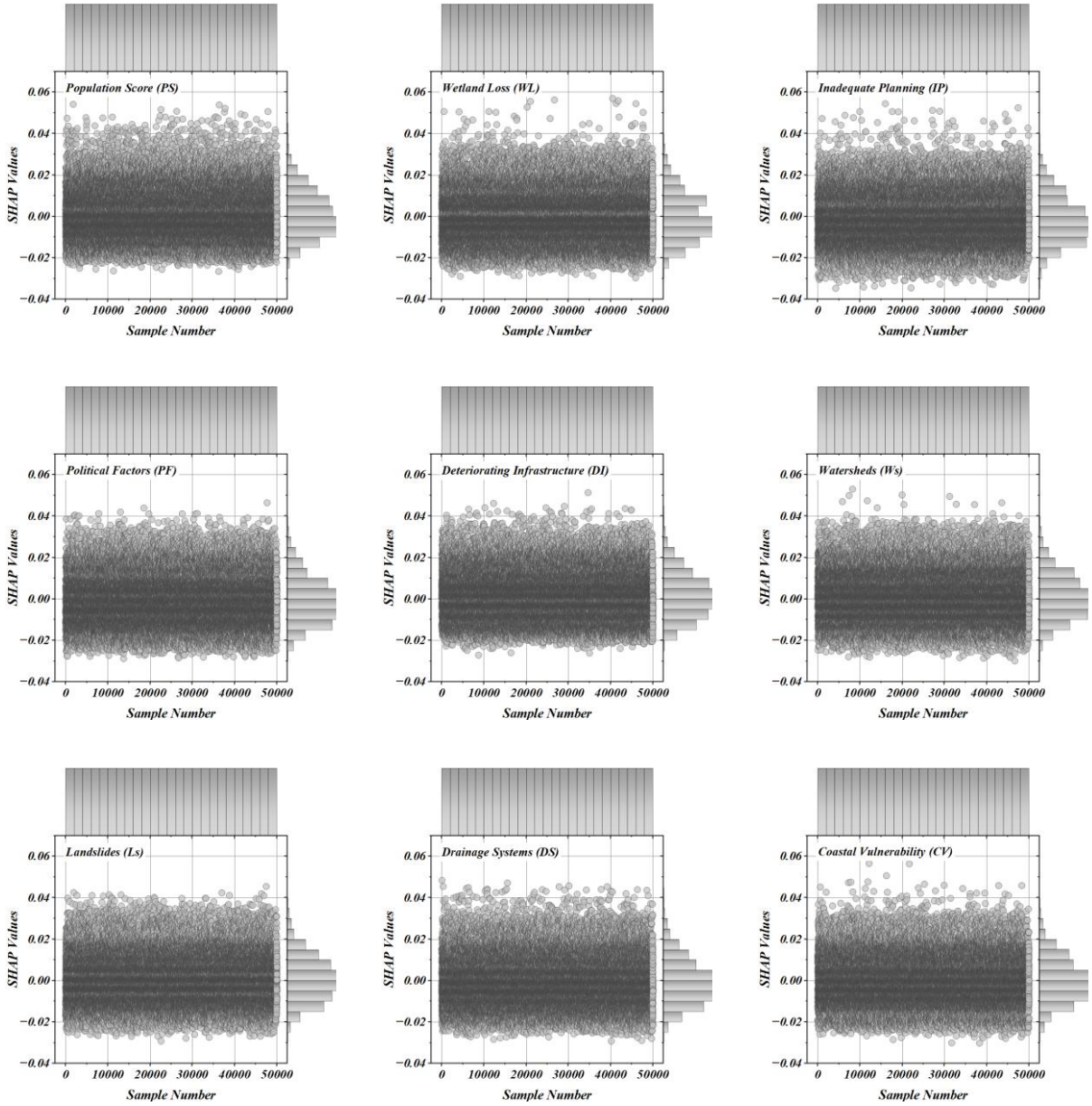
Table A6. Correlation of the variables within dataset concerning SVR model

DQ	*En*	*PF*	*Ls*	*AP*	*RM*	*Variable*	*FP*	*Ws*	*Ur*
0.008	0.009	-0.006	-0.004	0.004	1	***RM***	0.224	-0.009	0.005
0.002	-0.007	-0.001	0.004	1		***AP***	0.229	0.005	0.004
0.000	-0.003	0.004	1			***Ls***	0.223	0.009	-0.001
0.002	-0.008	1				***PF***	0.225	-0.003	-0.002
0.002	1					***En***	0.223	0.004	-0.001
1						***DQ***	0.227	0.004	-0.001
						CV	0.221	0.003	0.007
						DF	0.227	0.005	0.000
						TD	0.227	0.001	0.002
						IP	0.229	0.005	0.000
						Ur	0.225	0.000	0.001
						IDP	0.229	0.000	-0.010
						CC	0.220	0.002	-0.011
						PS	0.218	0.004	-0.011
						DS	0.222	0.004	-0.001
						Ws	0.218	-0.003	-0.005
						Si	0.221	-0.003	1
						MI	0.228	1	
						FP	1		

Si	*Ws*	*DS*	*PS*	*CC*	*IDP*	*Ur*	*IP*	*TD*	*DF*	*CV*
0.000	0.000	0.002	-0.001	0.007	0.006	-0.010	0.000	0.002	0.005	0.000
-0.004	0.004	0.001	0.002	-0.003	0.003	-0.001	-0.002	-0.005	0.003	-0.004
-0.002	0.004	0.004	0.002	0.001	-0.002	-0.001	-0.009	-0.002	0.005	0.003
-0.002	-0.003	0.001	0.007	-0.004	0.001	-0.002	-0.004	0.008	0.001	-0.003
0.000	0.004	-0.006	0.004	0.000	0.001	-0.011	0.001	-0.007	-0.004	-0.007
0.003	0.005	0.004	0.007	-0.003	-0.001	0.000	0.002	-0.004	-0.001	-0.010
-0.004	0.002	0.008	0.002	-0.001	0.001	-0.001	-0.010	-0.001	-0.003	1
-0.001	0.002	-0.001	0.000	0.001	0.002	-0.011	-0.001	0.002	1	
0.011	0.005	-0.002	0.006	-0.002	0.001	0.000	0.005	1		
0.012	0.009	-0.008	-0.002	0.000	0.004	-0.001	1			
-0.001	-0.003	-0.005	0.002	0.007	0.001	1				
0.001	0.000	0.004	-0.004	-0.005	1					
0.002	0.003	-0.008	-0.008	1						
-0.002	0.001	0.000	1							
-0.008	-0.003	1								
0.004	1									
1										

MI	0.002	0.003	-0.002	0.005	-0.004	0.003	-0.005	-0.005	-0.003	0.003	0.005	0.002	0.006	-0.004	0.000	-0.009	-0.003	1	
FP	0.229	0.222	0.223	0.225	0.218	0.227	0.215	0.220	0.229	0.223	0.221	0.225	0.221	0.227	0.218	0.228	0.227	0.224	1

Appendix 3



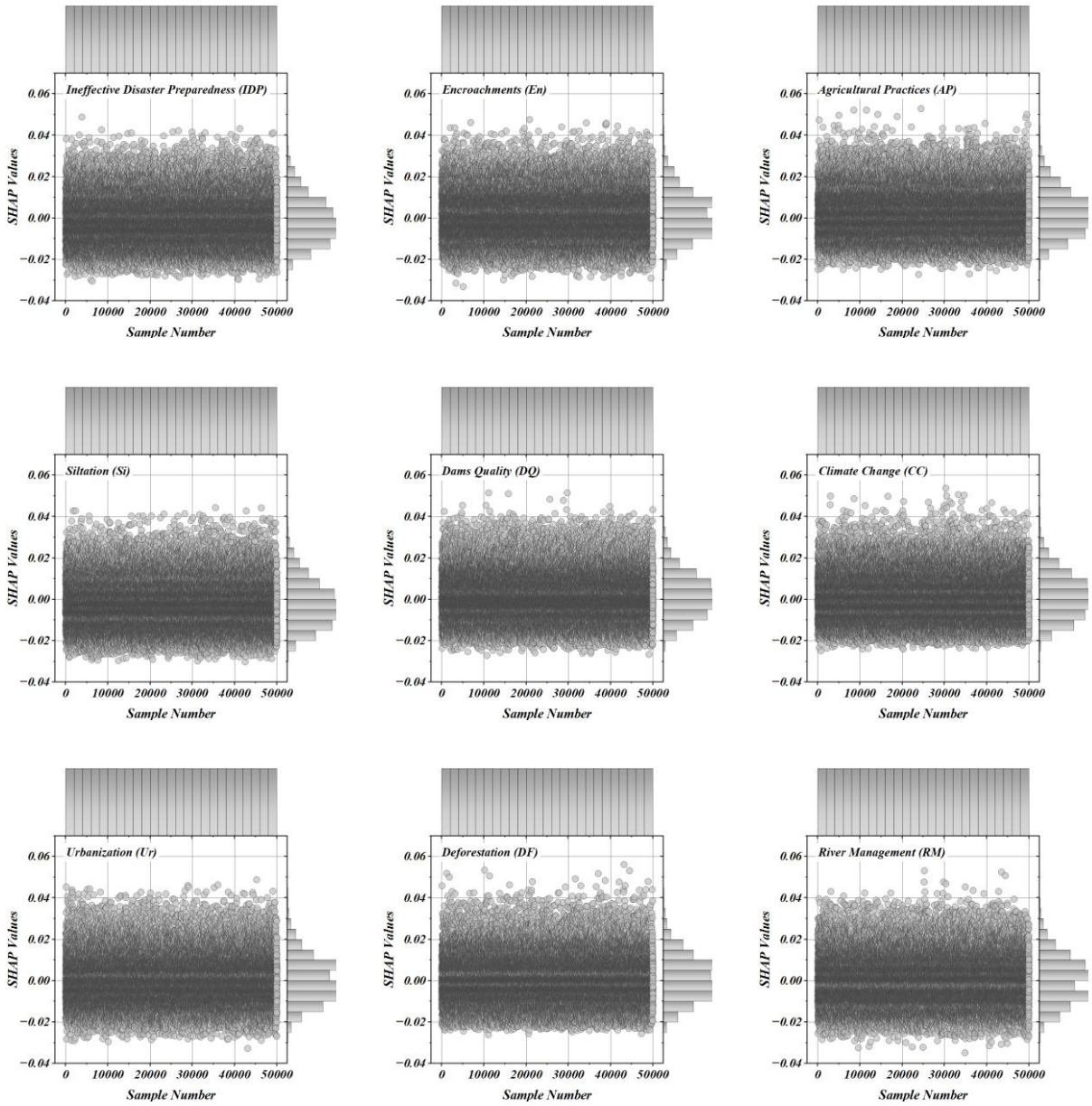


Fig. A1 SHAP sensitivity results of the variables within the dataset

References

Ahmad A, Farooq F, Niewiadomski P, et al (2021) Prediction of compressive strength of fly ash based concrete using individual and ensemble algorithm. Materials 14:794

Alin A (2010) Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics 2:370–374. https://doi.org/10.1002/wics.84

Alshammari E, Rahman AA, Rainis R, et al (2023) The impacts of land use changes in urban hydrology, runoff and flooding: a review. Current Urban Studies 11:120–141

Andreadis KM, Wing OEJ, Colven E, et al (2022) Urbanizing the floodplain: global changes of imperviousness in flood-prone areas. Environmental Research Letters 17:104024

Baig MA, Xiong D, Rahman M, et al (2022) How do multiple kernel functions in machine learning algorithms improve precision in flood probability mapping? Natural Hazards 113:1543–1562

Bazai NA, Alam M, Cui P, et al (2025) Dynamics and impacts of monsoon-induced geological hazards: a 2022 flood study along the Swat River in Pakistan. Natural Hazards and Earth System Sciences 25:1071–1093

Brownlee J (2020) Recursive feature elimination (RFE) for feature selection in Python. Machine Learning Mastery 25:

Carvalho P, Spataru C (2023) Gaps in the governance of floods, droughts, and heatwaves in the United Kingdom. Front Earth Sci (Lausanne) 11:1124166

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp 785–794

Contributors P (2024) perMetrics documentation. https://permetrics.readthedocs.io/en/latest/index.html#id8

Demissie Z, Rimal P, Seyoum WM, et al (2024) Flood susceptibility mapping: Integrating machine learning and GIS for enhanced risk assessment. Applied Computing and Geosciences 23:100183. https://doi.org/10.1016/j.acags.2024.100183

Der Sarkissian R, Al Sayah MJ, Abdallah C, et al (2022) Land Use Planning to Reduce Flood Risk: Opportunities, Challenges and Uncertainties in Developing Countries. Sensors 22

Erdal HI (2013) Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. Eng Appl Artif Intell 26:1689–1697

Fatica S, Kátay G, Rancan M (2024) Floods and firms: vulnerabilities and resilience to natural disasters in Europe. Available at SSRN 4796097

Ghasemi M, Zare M, Zahedi A, et al (2024) Optimization based on performance of lungs in body: Lungs performance-based optimization (LPO). Comput Methods Appl Mech Eng 419:116582. https://doi.org/https://doi.org/10.1016/j.cma.2023.116582

Granados GE, Miorelli R, Gatti F, Clouteau D (2023) A deep learning framework for efficient global sensitivity analysis and shap values calculations applied to eddy current testing problems. In: 50th annual review of progress in quantitative nondestructive evaluation. American Society of Mechanical Engineers, p V001T10A001

Han F, Yu J, Zhou G, et al (2024) Projected urban flood risk assessment under climate change and urbanization based on an optimized multi-scale geographically weighted regression. Sustain Cities Soc 112:105642

Hrast Essenfelder A, Toreti A, Seguini L (2025) Expert-driven explainable artificial intelligence models can detect multiple climate hazards relevant for agriculture. Commun Earth Environ 6:207. https://doi.org/10.1038/s43247-024-01987-3

Ibrahim A, Salifu A-H, Peprah C (2023) Does governance matter when disaster looms? Zooming into proactive institutional measures for flood risk management. International Journal of Disaster Risk Reduction 97:104021. https://doi.org/10.1016/j.ijdrr.2023.104021

Islam R, Chowdhury P (2024) Local-scale flash flood susceptibility assessment in northeastern Bangladesh using machine learning algorithms. Environmental Challenges 14:100833. https://doi.org/10.1016/j.envc.2023.100833

Khalid N (2024) Flood Prediction Dataset. https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset/versions/1. Accessed 30 Jul 2026

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. Montreal, Canada, pp 1137–1145

Li D-Y, Xu W, Zhao H, Chen R-Q (2009) A SVR based forecasting approach for real estate price prediction. In: 2009 International conference on machine learning and cybernetics. IEEE, pp 970–974

Li S, Tang Y, Zhao Y, et al (2025) Assessing future flood risk using remote sensing and explainable machine learning: A case study in the Beijing-Tianjin-Hebei region. Remote Sens Appl 40:101742. https://doi.org/10.1016/j.rsase.2025.101742

Moises DRJ, Kgabi NA, Kunguma O (2024) Policy implementation: Assessing institutional coordination and communication for flood warning in Namibia. Jàmbá: Journal of Disaster Risk Studies 16:1–10

Monachese AP, Gómez-Villarino MT, López-Santiago J, et al (2025) Challenges and Innovations in Urban Drainage Systems: Sustainable Drainage Systems Focus. Water (Basel). 17

Nguyen HD, Nguyen Q-H, Dang DK, et al (2024) A novel flood risk management approach based on future climate and land use change scenarios. Science of The Total Environment 921:171204

O'brien RM (2007) A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6

Potić I, Mihajlović LM, Šimunić V, et al (2022) Deforestation as a cause of increased surface runoff in the catchment: Remote Sensing and Swat Approach—A case study of southern Serbia. Front Environ Sci 10:896404

Pradhan B, Lee S, Dikshit A, Kim H (2023) Spatial flood susceptibility mapping using an explainable artificial intelligence (XAI) model. Geoscience Frontiers 14:101625. https://doi.org/10.1016/j.gsf.2023.101625

Pramanik MK, Talukdar S, Manna H, et al (2026) Flood Susceptibility and Mitigation in Tropical Cities Using XAI and 'Knowledge-Infused Transformer-based Neural Network' Model. Remote Sens Appl 102180. https://doi.org/10.1016/j.rsase.2026.102180

Rosenzweig B, Montalto FA, Orton P, et al (2024) NPCC4: Climate change and New York City's flood risk. Wiley Online Library

Sharma S, Tomar V, Arora S, Gupta N (2024) Flood Risk Assessment and Interpretation Using Explainable AI

Singha C, Rana VK, Pham QB, et al (2024) Integrating machine learning and geospatial data analysis for comprehensive flood hazard assessment. Environmental Science and Pollution Research 31:48497–48522

Sseguya F, Jun K-S (2024) Deep Learning Ensemble for Flood Probability Analysis. Water (Basel) 16:3092. https://doi.org/10.3390/w16213092

Taromideh F, Fazloula R, Choubin B, et al (2024) Ensemble Machine Learning for Urban Flood Hazard Assessment. In: 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, pp 525–530

Vojtek M, Janizadeh S, Vojteková J (2023) Riverine flood potential assessment using metaheuristic hybrid machine learning algorithms. J Flood Risk Manag 16:e12905. https://doi.org/10.1111/jfr3.12905

Wang J, Sanderson J, Iqbal S, Woo WL (2025) Accelerated and Interpretable Flood Susceptibility Mapping Through Explainable Deep Learning with Hydrological Prior Knowledge. Remote Sens (Basel) 17:1540

Wu P, Wang T, Wang Z, et al (2025) Impact of Drainage Network Structure on Urban Inundation Within a Coupled Hydrodynamic Model. Water (Basel). 17

Yang W, Zhang J, Hua P, Krebs P (2023) Global framework for flood risk management under climate change and urbanization. The Innovation Geoscience 1:100009. https://doi.org/10.59717/j.xinn-geo.2023.100009

Yuwono BD, Sabri LM, Wijaya AP, Awaluddin M (2024) Assessment of Flood Risk Induced by Land Subsidence Using Machine Learning. Indones J Geogr 56:376–386

Zhou D, Tawiah VK, Alessa N (2024) The impact of corruption on climate risk–An international evidence. Environ Dev 52:101059