computers

MDPI

Article

A Hybrid Ensemble for Early Flood Risk Forecasting Using Multimodal Spatiotemporal Data

Assylzat Slanbekova [1], Madi Akhmetzhanov [2,*], Leyla Fazylova [3], Shynar Turmaganbetova [4,*], Dinara Ussipbekova [5], Moldir Yessenova [1,*], Almira Mukhamejanova [6], Zhana-Gul Yessendauletova [3] and Zhanat Manbetova [7]

[1] Department of Information Systems, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; slanbekovaae@mail.ru
[2] Department of Applied Informatics and Programming, M. Kh. Dulaty Taraz University, Taraz 080000, Kazakhstan
[3] Department of Applied Mathematics and Computer Science, Karaganda National Research University Named After E.A. Buketov, Karaganda 100000, Kazakhstan; leyla.fazilova@mail.ru (L.F.); esendauletova81@mail.ru (Z.-G.Y.)
[4] Academy of Public Administration Under the President of the Republic of Kazakhstan, Astana 010000, Kazakhstan
[5] Department of Information and Communication Technologies, Faculty of International, Non-Profit Joint Stock Company S. Asfendiyarov Kazakh National Medical University, Almaty 010002, Kazakhstan; usipbekova.d@kaznmu.kz
[6] Department of Telecommunication Engineering, Almaty University of Power Engineering and Telecommunications Named After Gumarbek Daukeyev, Almaty 010002, Kazakhstan; a.mukhamejanova@aues.kz
[7] Educational Program Group, Department of Radio Engineering, Electronics, and Telecommunications, Institute of Power Engineering, S. Seifullin Kazakh Agrotechnical Research University, Astana 010000, Kazakhstan; zmanbetova@inbox.ru
* Correspondence: madiktz@gmail.com (M.A.); shynar.qurmangali@gmail.com (S.T.); moldir_11.92@mail.ru (M.Y.)

Abstract

This study presents a leakage-proof hybrid machine learning framework for early flood risk forecasting using multimodal spatiotemporal tabular data. An event-based dataset covering Kazakhstan from 2001 to 2021 was created by integrating topographic characteristics, hydrometeorological variables, long-term surface water dynamics, and remotely sensed spectral indices. To ensure realistic assessment, only pre-event observations were used, and event-based temporal data separation was employed to prevent leakage between the training and test subsets. The proposed Remote Sensing Adaptive Linear Opinion Pool Machine Learning (RS-ALOP-ML) framework combines multiple logistic regression experts with different regularization strengths through an Adaptive Linear Opinion Pool (ALOP) probabilistic fusion strategy, thereby preserving interpretability while improving forecasting robustness. The proposed framework was evaluated for four independent forecast horizons (T + 1, T + 7, T + 14, and T + 30 days) and compared with traditional machine learning algorithms and state-of-the-art tabular deep learning models, including Random Forest, ExtraTrees, XGBoost, LightGBM, CatBoost, MLP, FT-Transformer, TabNet, and Process Wide&Deep. Experimental results demonstrate that the proposed hybrid approach consistently achieves competitive or superior forecasting performance while maintaining computational efficiency and transparent probabilistic outputs. The study highlights that carefully designed hybrid machine learning architectures, combined with leakage-safe evaluation protocols, provide a robust foundation for multimodal environmental forecasting and decision support applications.

Academic Editor: Paolo Bellavista

Received: 3 July 2026
Revised: 3 September 2026
Accepted: 7 September 2026
Published: 10 September 2026

Keywords: hybrid machine learning; interpretable artificial intelligence; multimodal data; flood prediction; spatiotemporal analysis; leakage-safe evaluation; tabular learning

1. Introduction

Floods are among the most widespread and destructive natural hazards, causing significant economic losses, environmental damage, and threats to human life [1–3]. In countries with pronounced climate variability and complex hydrological systems, such as Kazakhstan, flood processes exhibit high spatiotemporal heterogeneity, which significantly complicates the task of early forecasting [4]. Traditional studies primarily focus on retrospective mapping of flood zones; however, this approach is not sufficiently effective for operational risk management, where forecasting before a hazardous event occurs is key [5]. Modern developments in remote sensing technologies, geographic information systems, and data analysis methods make it possible to integrate diverse sources of information to solve flood forecasting problems [6,7]. In particular, the use of satellite flood catalogs, long-term water dynamics products, hydrometeorological reanalyses, and relief and soil characteristics opens up new possibilities for constructing spatiotemporal risk models [8–10]. Multimodal data enable consideration of both static territorial features and dynamic pre-event factors, including precipitation accumulation, changes in soil moisture, and vegetation cover [11]. Despite significant progress, several important challenges remain in developing robust early flood forecasting systems. Existing research has not yet established whether leak-safe datasets based solely on pre-event observations are sufficient for robust predictive modeling without reliance on post-event information [12]. Furthermore, the performance and stability of multimodal predictors, including topographic features, long-term surface water dynamics, hydrometeorological variables, and remotely sensed spectral indices, across multiple forecast horizons remain poorly understood [13]. Moreover, although modern tabular deep learning methods have recently attracted considerable attention, it is still unclear whether they are superior to consistently interpretable machine learning approaches for multimodal spatiotemporal flood forecasting problems [14]. Finally, despite the growing interest in explainable artificial intelligence, limited attention has been paid to assessing whether the most influential predictors identified by forecasting models correspond to physically significant hydrological processes, especially in data-scarce regions [15].

Kazakhstan represents a particularly relevant case study for early flood forecasting due to its pronounced hydroclimatic heterogeneity and the diversity of flood mechanisms observed across the country. The northern and central regions predominantly suffer from spring floods caused by snowmelt, while the eastern and southeastern mountainous regions experience floods caused by rainfall and snowmelt under complex topographic conditions. Furthermore, the relatively sparse hydrological observation network and limited availability of long-term event data make Kazakhstan a representative example of a data-poor region where operational flood forecasting remains particularly challenging. Therefore, developing reliable, leak-proof forecasting systems for Kazakhstan is of practical importance not only for disaster risk reduction at the national level but also for other regions facing similar environmental and data constraints.

The methodological novelty of this study does not lie in the individual use of horizon-specific predictors, logistic regression, or probabilistic ensemble learning, which already have precedents in the literature. Rather, it contributes to their coordinated integration within the proposed RS-ALOP-ML framework for flood risk forecasting across multiple horizons, taking leakage risk into account. RS-ALOP-ML (Remote Sensing Adaptive Linear

Opinion Pool Machine Learning) is a horizon-specific probabilistic ensemble architecture that combines logistic regression experts with varying degrees of regularization via adaptive linear opinion pooling (ALOP). Conceptually, each expert processes the same leak-safe multimodal feature space but applies different regularization strengths, producing complementary probability estimates that are subsequently combined into a single flood risk probability. Specifically, RS-ALOP-ML generates a homogeneous ensemble of logistic experts for each horizon. These experts use the same multimodal feature space and probabilistic formulation but systematically vary in their regularization strength. This regularization-based design ensures controlled model diversity without introducing heterogeneous training models, while maintaining coefficient-level interpretability. The outputs of the probability experts are subsequently combined using adaptive linear pooling rather than majority voting, fixed averaging, or a separately trained stacked classifier. The resulting architecture thus combines three elements within a single forecasting procedure: an event-horizon- and time-specific assessment, controlled diversity generated within a single interpretable family of models, and adaptive probability-level fusion. The novelty claimed here lies precisely in this coordinated architecture and its application to multimodal early flood risk forecasting, rather than in examining each of these methods separately.

The main findings of this study are as follows:

1. A coordinated RS-ALOP-ML forecasting architecture that generates controlled diversity through regularization-specific homogeneous logistic experts and combines their results using adaptive linear opinion fusion over independently constructed forecast horizons that eliminate data leakage.
2. A comprehensive evaluation protocol combining time-sampling validation, bootstrapped confidence intervals, probability calibration, SHapley Additive exPlanations (SHAP)-based interpretation, and transferability analysis for operational flood hazard forecasting.
3. A physically interpretable probabilistic system that estimates flood hazard probability based on multimodal geospatial and hydrometeorological information under realistic early warning conditions.

The article is structured as follows. Section 2 presents a literature review on modern approaches to flood forecasting and multimodal spatiotemporal modeling. Section 3 describes the data used and the research methodology, including the formation of a multimodal event-driven dataset and the construction of a leakage-safe pipeline. Section 4 presents the results of the experimental analysis and a comparative evaluation of the models over different forecast horizons. Section 5 discusses the results, their interpretation, and the study's limitations. The conclusion formulates the main findings and outlines areas for further research.

2. Related Works

In recent years, there has been a steady increase in interest in applying machine learning methods to flood forecasting and hydrological risk assessment. Several studies demonstrate the effectiveness of integrating hydrometeorological and satellite data to improve forecast accuracy and model robustness [16]. In particular, it has been shown that combining remote sensing with ML algorithms improves the spatial generalizability of models and reduces dependence on ground-based observations [17,18]. Ensemble methods, including Random Forest and gradient boosting, which demonstrate high robustness to noise and data heterogeneity [19,20], are actively used in flood susceptibility mapping tasks. At the same time, modern studies emphasize that model quality is determined not only by the algorithm but also by the proper formation of the training sample and the validation procedure [21]. Special attention is paid to the use of hydrometeorological reanalyses

and long-term time series. It has been shown that ERA5 data and similar sources enable the formation of physically consistent features that reflect precipitation accumulation, the temperature regime, and snow cover conditions, significantly improving the quality of flood forecasting [22,23].

In recent years, the application of deep neural network architectures to tabular and spatiotemporal data has been actively explored. Models such as FT-Transformer, TabNet, and hybrid Wide&Deep approaches demonstrate potential for integrating heterogeneous features [24,25]. However, comparative studies show that for tabular environmental data, classical machine learning methods often remain more robust and interpretable [26]. The field of interpretable machine learning has continued to develop. Explainability methods, including SHAP, enable analysis of feature contributions and verification of model physical validity. This is especially important for natural hazard forecasting, where the model must be based on real hydrological and climatic factors rather than random correlations [27].

A number of contemporary studies consider spatiotemporal models that account for the dynamics of hydrological processes in both time and space. Such approaches include recurrent neural networks (RNNs), which process sequential data by propagating information across time steps; long short-term memory (LSTM) networks, a specialized type of RNN designed to account for long-term temporal dependencies; and temporal convolutional networks (TCNs), which use convolutional operations along the time dimension to model sequential patterns. In flood forecasting, these architectures can account for the cumulative and time-dependent effects of pre-event factors such as precipitation, soil moisture, and temperature [28,29].

A separate area of research involves the use of geospatial methods and GIS analysis for flood hazard assessment. Such studies widely employ multicriteria modeling (MCDM) methods, including Analytic Hierarchy Process (AHP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and their hybrid combinations with machine learning [30,31]. Despite their high interpretability, such approaches often rely on expert judgment and may not always account for complex nonlinear relationships among factors.

Probabilistic and risk-based modeling is also developing, in which forecasting is considered not as a binary decision, but as an assessment of the probability of flood occurrence. Such approaches allow for explicit representation of forecast uncertainty and facilitate the integration of model results into risk-based decision support systems [32]. In this context, the calibration and reliability of probabilistic forecasts are particularly important. The Brier score quantifies the root-mean-square difference between predicted probabilities and observed binary outcomes, with lower values indicating more accurate probabilistic forecasts, while reliability diagrams compare predicted probabilities with observed event frequencies within probability intervals to visually assess calibration [33]. Thus, these additional measures help determine whether predicted flood probabilities can be interpreted as meaningful estimates of the actual probability of an event.

Despite significant progress in machine learning-based flood forecasting, relatively few studies simultaneously integrate leakage safety assessment, multimodal data fusion, horizon-specific forecasting, and probabilistic model interpretation within a single early warning system. Existing studies tend to consider these aspects separately, focusing on leakage prevention, spatiotemporal modeling, ensemble forecasting, or model interpretability. Consequently, there remains a need for forecasting systems that jointly provide realistic out-of-sample assessment, robust probabilistic forecasting, and interpretable decision support in operational early warning settings [34–37]. This methodological gap motivates the integrated framework explored in this study. More broadly, the rapid advancement of artificial intelligence has shifted research attention from maximizing predictive performance alone to developing robust, interpretable, and decision-support-oriented AI systems

capable of operating in complex real-world settings. This shift has facilitated the adoption of explainable and robust machine learning approaches in various application areas, including environmental monitoring and disaster prediction [38].

Although probabilistic ensemble learning and linear opinion pooling have been explored in a number of machine learning applications, their use in environmental forecasting remains limited. Existing flood forecasting studies primarily employ heterogeneous ensembles based on Random Forests, Gradient Boosting, or deep neural networks, while coordinated ensembles of homogeneous logistic experts pooled using adaptive linear opinion pooling have been rarely explored. Thus, an analysis of existing research shows that further development of flood forecasting methods requires an integrated approach combining multimodal data, a well-defined problem statement, reproducible methodologies, and interpretable models capable of operating under real-world constraints and uncertainty.

3. Materials and Methods

This section presents the research methodology, including a description of the data used, the procedures for their preparation, and approaches to constructing models for early flood hazard forecasting. Unlike traditional retrospective approaches, this study employs a rigorous forecasting framework that relies exclusively on pre-event information. Particular attention is paid to the formation of a multimodal event-oriented dataset combining geospatial, hydrometeorological, and satellite features, as well as the development of a leakage-safe pipeline to ensure the correctness of model evaluation and the reproducibility of results. The proposed methodology aims to build robust and interpretable models capable of accounting for the complex spatiotemporal nature of flood processes.

In the proposed framework, the forecasting problem is formulated as a probabilistic assessment of flood hazard rather than a deterministic forecast of future hydrological or meteorological variables. For each forecast horizon (T + 1, T + 7, T + 14, and T + 30 days), the RS-ALOP-ML model estimates the probability of a flood occurrence using only pre-event geospatial and hydrometeorological information available up to the corresponding forecast date. Consequently, each forecast horizon represents an independent flood hazard forecast scenario constructed based on pre-event information, and the reported performance reflects the model's ability to estimate the probability of a flood hazard under various pre-event information conditions.

3.1. Dataset

This study utilizes a purpose-built multimodal event-based dataset for early spatiotemporal flood hazard forecasting. Unlike existing benchmark datasets, it was developed using leakage-free design principles to support predictive modeling under conditions as close as possible to real-world early warning scenarios. Structurally, the dataset consists of two interconnected subsets: a full integrated dataset covering the period 2001–2021, and a forecast-ready subset containing only pre-event observations available at four forecast horizons (T + 1, T + 7, T + 14, and T + 30 days). This design eliminates the use of post-event information and ensures that the problem is formulated as a true forecasting problem, not a retrospective event recognition problem. The first group includes flood labels obtained primarily from the Global Flood Database version 1 (2000–2018), available through Google Earth Engine as GLOBAL_FLOOD_DB/MODIS_EVENTS/V1. The dataset was derived from MODIS Terra and Aqua observations and provides spatial and temporal information on documented major floods at a native flood classification resolution of 250 m. In this study, it was used to determine the dates, locations, and spatial extent of verified floods. The methodology underlying the Global Flood Database is described by Tellman et al. [39]. To extend the temporal coverage beyond the official database period, four independently

collected floods for 2019–2021, verified using publicly available documentary sources, were additionally included. The second group includes static and quasi-static geospatial variables describing flood susceptibility, including long-term surface water characteristics from the JRC global surface water dataset, topographic features derived from SRTM, and soil texture information. The third group contains hydrometeorological variables obtained from ERA5-Land, including antecedent precipitation, air temperature, soil moisture, dew point, and snow depth, aggregated over multiple pre-event time windows (3, 7, 14, and 30 days). Although individual floods were not explicitly classified by their generation mechanism, these predictor variables were specifically chosen to characterize the dominant hydrometeorological processes responsible for flood occurrence in Kazakhstan. Consequently, the proposed framework is capable of representing snowmelt-induced floods, rainfall-induced floods, and mixed hydrometeorological flood conditions using physically meaningful pre-event predictors rather than predefined event labels. The fourth group includes MODIS-derived spectral indices (NDVI, EVI, NDWI, MNDWI, NDMI, and BSI), which characterize land surface conditions prior to flood occurrence.

The hydrometeorological predictors were obtained from the ERA5-Land reanalysis, whose source fields are presented at an hourly temporal resolution on a regular spatial grid of $0.1^\circ \times 0.1^\circ$ (approximately 9 km at the equator). In this study, the data were extracted from the daily aggregated product ECMWF/ERA5_LAND/DAILY_AGGR with a nominal pixel size of 11,132 m. The hourly source fields were not used directly as model input; instead, precipitation, air temperature, dewpoint temperature, snow depth, and soil moisture data were aggregated for strictly pre-event periods of 3 to 30 days, depending on the predictor and forecast horizon. Because the source datasets differed significantly in their original spatial resolution, they were integrated using a coordinate-based feature extraction procedure rather than by resampling all raster products to a common grid. Each event-based observation was associated with geographic coordinates (longitude and latitude), which served as a common spatial reference for extracting predictor values from ERA5-Land, MODIS, SRTM, JRC Global Surface Water, and OpenLandMap. This procedure preserved the original spatial resolution of each source while simultaneously establishing spatial matching at the observation level. Temporally varying predictors were additionally aligned with the corresponding forecast date and summarized only within predefined windows before the event, preventing the inclusion of information unavailable at the time of forecasting. The resulting source-specific characteristics were then combined into a single event-based tabular record for each observation. The original spatial and temporal resolutions of all data products, as well as the corresponding spatial and temporal matching procedures, are provided in Table S2.

Several limitations of the resulting dataset should be noted. First, since the Global Flood Database is primarily based on MODIS observations, it predominantly represents medium- and large-scale floods, while localized or short-term floods may be underrepresented, particularly in regions with complex topography or limited satellite coverage. This limitation was partially mitigated by incorporating locally selected flood data for 2019–2021, which expanded the temporal coverage and improved the completeness of event data. Second, the proposed framework integrates multiple geospatial and environmental datasets with different spatial resolutions, temporal frequencies, and data collection characteristics. Therefore, the resulting feature space should be interpreted as a consistent multimodal representation of flood risk rather than as a homogeneous source of observations. To minimize the impact of source heterogeneity, all variables were transformed into a single event-based tabular representation containing only information available before the forecast date. Furthermore, leakage-free preprocessing was consistently applied throughout the experimental process, including event-based temporal partitioning, imputing missing data only from

the training set, standardizing data only from the training set, and developing a model for each forecast horizon. Despite these measures, residual uncertainty associated with differences in spatial and temporal resolution cannot be completely eliminated and must be considered when transferring the proposed framework to highly localized flood forecasting applications. To ensure methodological rigor, all variables potentially containing leaky target information, including post-event characteristics and service identifiers, were removed before model development. The remaining observations were divided according to flood events to prevent records of the same event from appearing simultaneously in different data subsets. As a result, the final dataset represents a leak-safe, multimodal, event-based spatiotemporal benchmark that integrates geospatial, hydrometeorological, and remote sensing information for early flood hazard forecasting across multiple forecast horizons. A detailed summary of the input data sources, their spatial and temporal characteristics, potential limitations, and the corresponding mitigation strategies adopted in this study is provided in Table S1.

3.2. Methodology for Data Analysis and the Formation of a Leakage-Safe Predictive Pipeline

This section presents a methodology for data preparation and analysis implemented for early spatiotemporal flood hazard forecasting. Unlike standard approaches focused on retrospective event recognition, the proposed method is based on a rigorous forecasting framework that uses only pre-event features available at the time of forecast generation. Therefore, special attention is paid to constructing a leakage-safe pipeline that eliminates target information leakage at all stages of data processing, from observation filtering to the generation of model-ready feature matrices. This approach improves the reproducibility of experiments, the accuracy of model evaluation, and the comparability of results when analyzing various machine and deep learning algorithms.

In this study, data leakage refers to the unintentional use of information during model training or evaluation that would not have been available at the time the corresponding flood forecast was issued. Such leakage can occur, for example, when post-event observational data, variables derived from future data, or information about the same flood event are used for both training and evaluation, leading to overly optimistic performance estimates. Therefore, a leakage-proof dataset contains only information about predictors available at or before the time of the corresponding forecast, and is partitioned so that information from evaluation events cannot leak into the model training process. The overall workflow of the proposed flood forecasting system is shown in Figure 1. The methodology consists of six sequential steps, starting with the collection of heterogeneous geospatial, hydrometeorological, topographic, and remote sensing data and integrating them into an event-based tabular representation. Next, a leakage-proof forecasting setup is applied by excluding information related to flood onset or post-flood conditions and constructing four independent forecast horizons (T + 1, T + 7, T + 14, and T + 30 days). Model development follows a decoupled validation strategy to ensure that observations of the same flood event are not used in different data partitions together with an independent time interval including events from 2019 to 2021. The proposed RS-ALOP-ML framework is subsequently compared with traditional machine and deep learning models. Finally, model performance is evaluated using discrimination, classification, and probabilistic metrics, complemented by ablation analysis and SHAP-based interpretation. This end-to-end workflow provides a consistent framework for leak-proof model development, temporal validation, comparative evaluation, and interpretation.

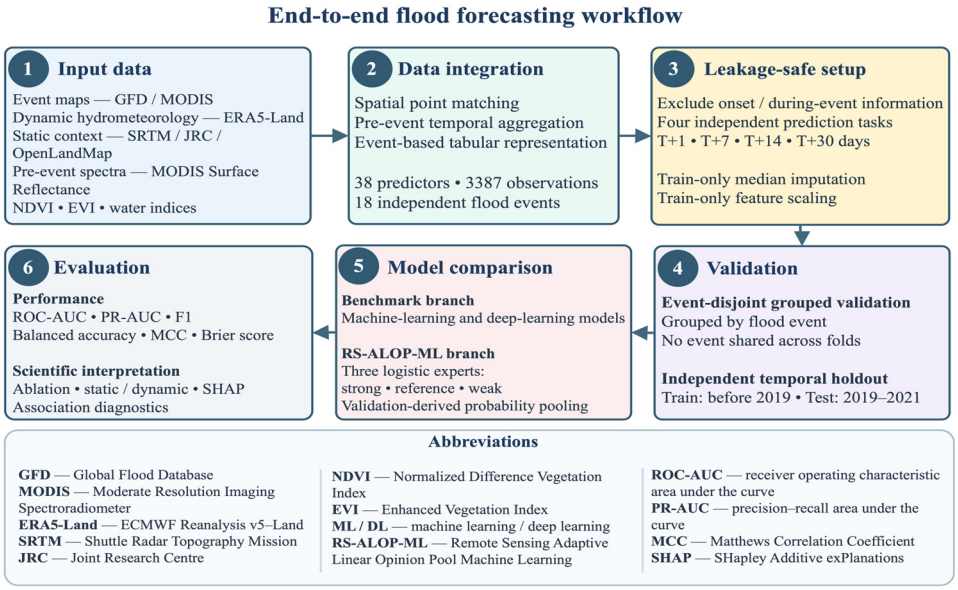

Figure 1. An end-to-end flow chart of the flood risk forecasting process.

The bottom block specifies the target variable y, considered as a binary flood hazard label, and indicates the forecast horizons $h \in \{1, 7, 14, 30\}$ days. The right-hand side illustrates the leakage-safe preprocessing sequence: from selecting only forecast-correct pre-event rows to constructing model-ready feature matrices. The scheme includes filtering forecast-ready observations, role-based variable mapping, cyclic seasonality coding, forming independent cohorts by horizons, event-safe sample partitioning, train-only gap imputation, and train-only standardization. The data analysis in this study was organized as a leakage-safe spatiotemporal pipeline for early flood hazard forecasting. The original table was considered not as a conventional static set of observations, but as an event-oriented set of rows, where each record corresponds to a specific spatial point, a specific event, and a specific forecast reference date (1):

$$\mathcal{D}_0 = \left\{ \left(\mathrm{x}_i^{raw}, y_i, e_i, h_i, t_i^{ref} \right) \right\}_{i=1}^{N_0} \tag{1}$$

where x_i^{raw} is the original feature vector, $y_i \in \{0, 1\}$ is the binary target label, e_i is the event identifier, $h_i \in \{1, 7, 14, 30\}$ is the forecast horizon in days, and t_i^{ref} is the date for which the features are available. The target variable was defined as a binary flood hazard indicator (2):

$$y_i = \begin{cases} 1, & \text{if the line refers to a flood-prone sample,} \\ 0, & \text{if the line refers to a non-flood control sample} \end{cases} \tag{2}$$

In the first stage, only those rows that satisfied the predictive suitability condition were selected from the total array. This means that only pre-event observations for which all information was available no later than the reference date t_i^{ref} were retained in the table. Formally, the filtering can be written as (3):

$$\mathcal{D}_{fr} = \left\{ \left(\mathrm{x}_i^{raw}, y_i, e_i, h_i, t_i^{ref} \right) \in \mathcal{D}_0 \mid f_i^{ready} = 1 \right\} \tag{3}$$

where f_i^{ready} is the predictive suitability indicator that determines whether a given observation is suitable for use in forecasting based on the temporal availability of its predictor information. An observation is considered suitable if all predictor values used for forecast-

ing would have been available on or before the corresponding forecast date. Observations containing information that became available only after the forecast date are excluded to prevent temporary data leakage and maintain realistic early warning settings. After filtering, the roles of the variables were determined for each row: the predictor matrix X, the binary target y, the event identifier e, and the horizon h. The final feature vector included 38 variables and was formed as a union of four thematic blocks (4):

$$x_i = \left[x_i^{(geo)}, x_i^{(hydro)}, x_i^{(spec)}, x_i^{(time)}\right], x_i \in \mathbb{R}^{38} \quad (4)$$

where $x_i^{(geo)}$ are relief-geographical features, $x_i^{(hydro)}$ are hydrometeorological aggregates, $x_i^{(spec)}$ are spectral surface indices, and $x_i^{(time)}$ are temporal and qualitative features. The geographic block included longitude, latitude, altitude, slope, aspect, soil texture class, as well as characteristics of the long-term water history of the area. The hydrometeorological block included pre-event precipitation amounts and average values of temperature, dew point, snow depth, and soil moisture over several time windows. The spectral block reflected the state of the underlying surface through the NDVI, EVI, NDWI, MNDWI, NDMI, and BSI indices. The temporal block contained observation quality indicators and seasonal features. Before training, all fields that could directly or indirectly encode an already observed event, service markup, or sampling structure were excluded. Let $\mathcal{F}_{all}$ be the set of all available fields, and $\mathcal{L}$ be the set of leakage variables. Then the final set of allowed predictors was defined as (5):

$$\mathcal{F} = \mathcal{F}_{all} / \mathcal{L} \quad (5)$$

After this, each row was transformed into a safe prediction vector (6):

$$\tilde{x}_i = \Pi_{\mathcal{F}}(x_i^{raw}) \quad (6)$$

where $\Pi_{\mathcal{F}}$ is the projection operator onto the set of allowed features. Thus, only the information that could be used in a real forecast scenario was included in the model space. Since calendar seasonality is cyclical in nature, the "day of year" variable was not used in linear form. Instead, it was encoded using sine and cosine (7):

$$doy_{\sin,i} = \sin\left(2\pi \frac{d_i}{366}\right), doy_{\cos,i} = \cos\left(2\pi \frac{d_i}{366}\right) \quad (7)$$

where d_i is the day-of-the-year number for the date t_i^{ref}. This representation eliminates the artificial gap between the end of December and the beginning of January and allows for the seasonal periodicity of flood processes to be correctly taken into account. The overall forecast table was then divided into four independent subtasks based on early warning horizons: T + 1, T + 7, T + 14, and T + 30 days. A separate subset (8) was formed for each horizon:

$$\mathcal{D}^{(h)} = \left\{ \left(\tilde{x}_i, y_i, e_i\right) \;\middle|\; h_i = h \right\}, h \in \{1,\ 7,\ 14,\ 30\} \quad (8)$$

This approach avoided mixing short-term and medium-term forecasting modes and allowed us to evaluate the quality of the models separately for each time window. The data was split to prevent the same event from being included in both training and validation. For internal quality assessment, group cross-validation by event identifier was used (9):

$$\mathcal{E}_{train}^{(k)} \cap \mathcal{E}_{val}^{(k)} = \varnothing \quad (9)$$

where $\mathcal{E}_{train}^{(k)}$ and $\mathcal{E}_{val}^{(k)}$ are the sets of events in the training and validation portions of the k-th fold. This meant that the model was tested on new events, not on rows of an already known event. An external time holdout was additionally used, in which training was performed on earlier events and testing on later ones (10):

$$\mathcal{D}_{train}^{(h)} = \left\{ i : \text{year}(t_i^{ref}) < 2019 \right\}, \mathcal{D}_{test}^{(h)} = \left\{ i : \text{year}(t_i^{ref}) \geq 2019 \right\} \tag{10}$$

The temporal separation boundary was deliberately set at 2019, rather than chosen arbitrarily. Events recorded between 2001 and 2018 were used for model development, as this period provides sufficient historical observations for reliable parameter estimation, while events from 2019 to 2021 represent an independent, future-oriented evaluation period. This design accurately reflects the operational forecasting scenario, in which models are trained using historical flood data and then applied to unknown future events. Consequently, the chosen time interval allows for a realistic assessment of the model's generalization ability under chronological unfolding conditions, while maintaining a sufficient amount of training data. This scheme simulated a real-world scenario for the operational use of the model, where a forecast is based on historical data and tested in subsequent years. Traditional random percentage splitting was not adopted as the primary estimation protocol because observations from the same flood event are statistically dependent. Therefore, random splitting could assign observations from the same event to both the training and test sets, leading to information leakage and overly optimistic performance estimates. For this reason, event-based temporal sampling was chosen as the primary estimation strategy. However, random percentage splitting may be considered in future studies as an additional sensitivity analysis to further explore the model's robustness under alternative validation conditions. Missing values in numerical features were replaced with medians calculated strictly based on the training subset. For the j-th feature, this can be written as (11):

$$x_{ij}^{imp} = \begin{cases} x_{ij}, & \text{if the value is observed,} \\ \text{median}\left(x_{\cdot j}^{train}\right), & \text{if the value is missing} \end{cases} \tag{11}$$

Using median imputation only in the training set eliminated leakage of statistical characteristics from validation and testing. After imputation, scale-sensitive features were standardized to the training set parameters (12):

$$z_{ij} = \frac{x_{ij}^{imp} - \mu_j^{train}}{\sigma_j^{train}} \tag{12}$$

where μ_j^{train} and σ_j^{train} are the mean and standard deviation of the j-th feature, calculated only on the training data. These same parameters were then applied to the validation and test rows without re-evaluation. The result of the preprocessing stage was model-ready matrices (13):

$$\mathrm{X}_{scaled}^{(h)} \in \mathbb{R}^{N_h \times 38}, \mathrm{y}^{(h)} \in \left\{0,1\right\}^{N_h} \tag{13}$$

where N_h is the number of observations for horizon h. For each forecast horizon, a probabilistic risk assessment function was constructed (14):

$$\hat{p}^{(h)}(\mathrm{x}) = P(Y = 1 \mid \mathrm{x}, h) \tag{14}$$

where $\hat{p}^{(h)}(\mathrm{x})$ is the probability that an observation belongs to the flood hazard class at horizon h. In other words, each model solved a binary probabilistic forecasting problem using a 38-dimensional multimodal feature vector. When using multiple competing models,

the final solution was compared using a set of quality metrics. For threshold classification, the probabilistic forecast was transformed into a final label as follows (15):

$$\hat{y}_i^{(h)} = \mathbb{I}\left[\hat{p}^{(h)}(\mathrm{x}_i) \geq \tau^{(h)}\right] \tag{15}$$

where $\tau^{(h)}$ is the threshold for horizon h, and $\mathbb{I}[\cdot]$ is the indicator function. This allowed us to translate the probabilistic risk into an interpretable binary decision. Ablation analysis was used to test the actual contribution of feature blocks and the robustness of the best model. If $M(\mathcal{F})$ is the value of the selected metric on the full feature set, and $M(\mathcal{F}/g)$ is the same value after removing group g, then the contribution of this group was determined as (16):

$$\Delta_g = M(\mathcal{F}) - M(\mathcal{F}/g) \tag{16}$$

A positive Δ_g value meant that removing the corresponding feature block degrades the forecast quality, and therefore this block makes an informative contribution to the model. In a compact form, the entire data analysis framework can be represented as a sequential transformation (17):

$$\mathcal{D}_0 \to \mathcal{D}_{fr} \to \left\{\mathcal{D}^{(1)}, \mathcal{D}^{(7)}, \mathcal{D}^{(14)}, \mathcal{D}^{(30)}\right\} \to \mathrm{X}_{scaled}^{(h)}, \mathrm{y}^{(h)} \to \hat{p}^{(h)}(\mathrm{x}) \tag{17}$$

where each transition corresponds to a separate stage: filtering forecast-correct rows, removing leakage fields, generating horizons, train-only imputation, train-only standardization, and subsequent model training. Thus, the applied methodology constitutes a reproducible and academically sound data-preparation and analysis scheme for early spatiotemporal flood-hazard forecasting. Its fundamental feature is that all stages, from selecting pre-event rows to constructing model-ready feature matrices, were performed within the logic of a rigorous forecast formulation, without using information unavailable at the time of forecasting. This ensures both the scientific validity of the results and their practical suitability for early warning tasks. Below are tables reflecting the scientifically significant characteristics of the resulting dataset: its composition, quality, and gap structure. They allow us to capture the sample's initial properties before the modeling stage and provide a formal basis for the subsequent interpretation of classification and forecasting results. The separate presentation of these summaries is important because it addresses typical questions about the completeness of event coverage, class balance, the presence of duplicates, and the suitability of the feature space for leakage-safe analysis.

The final integrated dataset contains 3387 observations representing 18 unique flood events recorded between 2001 and 2021. The number of flood-related and non-flood-related observations is roughly balanced, indicating no significant class imbalance at the initial stage of dataset construction. The forecast-ready subset includes 2701 observations, maintaining the same set of 18 flood events and full temporal coverage, confirming that the transition to forecast formulation has not impacted event diversity. Additionally, to expand the temporal coverage beyond the official Global Flood Database, a locally compiled event registry consisting of four flood events was included.

Quality control confirmed the structural consistency of both the integrated and forecast-ready datasets. No duplicate observations were found in either subset, indicating that the data preparation pipeline successfully eliminated redundant records. The forecast-ready dataset consists exclusively of 2701 pre-event observations, ensuring that only pre-flood information was used for model development. In contrast, the full integrated dataset additionally contains 686 pre-event observations, which were intentionally excluded from the forecast subset to prevent leakage of target information. Furthermore, eight variables

with a high proportion of missing values were consistently identified in both datasets and subsequently processed using a leakage-proof preprocessing strategy.

Missing value analysis revealed that the largest proportion of missing data is concentrated in the OpenMeteo alternative variables, each of which exhibits a missing data fraction of 1.00 (100%), indicating that these variables are not available for the final operational forecasting pipeline. Among the remaining variables, only post-event observation descriptors, including observed_event_duration_days, observed_local_water_anomaly_proxy, observed_distance_to_flood_core, and observed_flood_fraction_local, exhibited a significant missing data fraction (79.75%). In contrast, all hydrometeorological predictors used for forecasting, including snow_depth_mean_7d, soil_moisture_mean_7d, soil_moisture_mean_14d, precip_sum_30d, dewpoint_mean_7d, and pre-event moisture derivatives, had a missing data fraction of 0%. These results confirm that the final set of predictor features consists exclusively of the complete variables available before the flood event, while variables describing the observed flood itself were intentionally excluded from the model development to prevent leakage of targeted information.

Figure 2 shows the distribution of the number of rows in the final data set by observation year. The abscissa axis represents years, and the ordinate axis represents the number of observations (Number of Rows) corresponding to each year. The analysis shows that the data volume varies significantly over time. The minimum values are observed in 2007 (20 rows), 2001 (40 rows), and 2004 (63 rows). Relatively low values are also characteristic of 2006 (97 rows) and 2016 (122 rows). The maximum data volume was recorded in 2005 (497 rows), as well as in 2003 (488 rows) and 2011 (445 rows). Significant values are observed in 2017 (360 rows), 2015 (345 rows), 2021 (275 rows), 2019 (250 rows), and 2020 (220 rows). For 2002, the sample size is 165 rows.

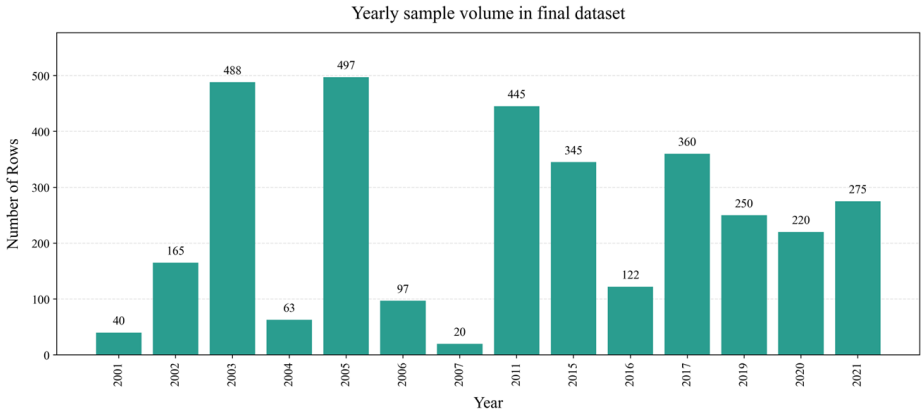

Figure 2. Annual distribution of the number of observations in the final data set.

The temporal distribution of flood events included in the final dataset is shown in Figure 3. A total of 18 independent flood events were identified between 2001 and 2021. Two events were recorded in 2003, 2004, 2005, and 2015, while the remaining years included one confirmed event each. Although the dataset contains a relatively limited number of independent events, it was designed to reflect a variety of flood conditions across Kazakhstan rather than to provide a complete historical inventory. The selected events represent the main flood regimes considered in this study, including snowmelt-induced floods in the northern and central regions, rainfall-induced floods in the eastern and southeastern mountainous areas, and mixed hydrometeorological flood conditions in other parts of the country. This selection of events provides geographic and hydrological diversity while maintaining an event-based structure suitable for robust predictive modelling and assessment.

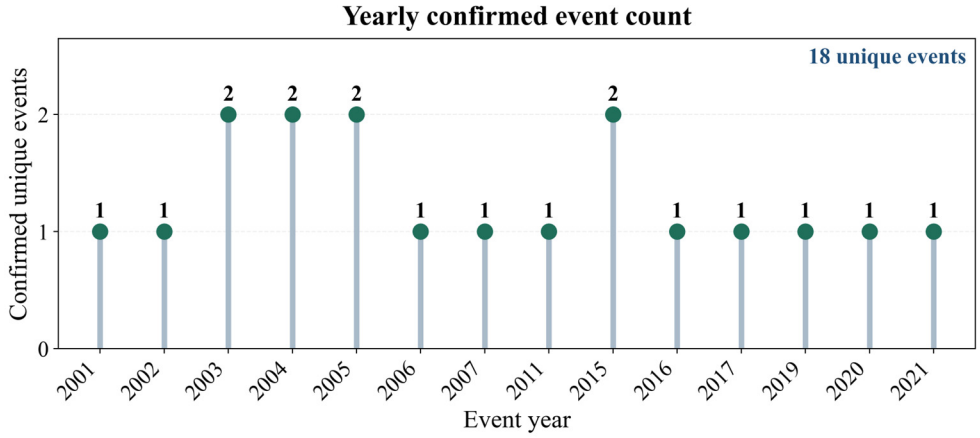

Figure 3. Annual distribution of the number of confirmed unique flood events.

Figure 4 shows the distribution of observations in the forecast-ready data subset across various pre-event time phases. The abscissa axis indicates the intervals before the event (pre_event_30d, pre_event_14d, pre_event_7d, pre_event_1d), and the ordinate axis shows the number of corresponding observations (Number of Rows). The analysis shows that the data volume is distributed relatively evenly across all time phases. The largest number of observations falls within the pre_event_30d interval—714 rows, while the minimum value is recorded for pre_event_14d-653 rows. For the pre_event_7d and pre_event_1d phases, the values are 655 and 679 rows, respectively. Minor differences between the values indicate a balanced sample across the forecast time horizons. This is an important factor for the correct training of models, since it eliminates bias towards one-time interval and ensures comparability of results when forecasting at different horizons (T + 1, T + 7, T + 14, and T + 30 days).

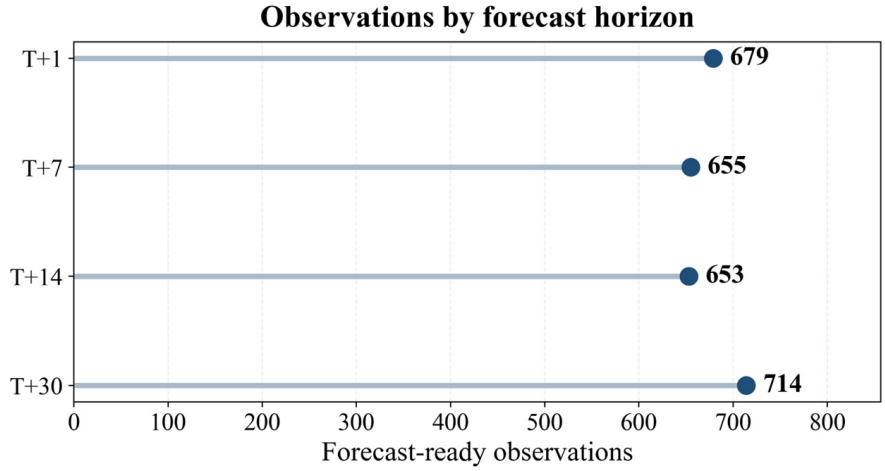

Figure 4. Distribution of the number of observations by time phases in the forecast-ready sample.

Figure 5 shows the distribution of classes in the full and forecast-ready datasets. In the full dataset, the non-flood class comprises 1736 rows, and the flood class comprises 1651 rows. In the forecast-ready subset, a similar ratio is maintained: 1380 rows belong to the non-flood class, and 1321 rows belong to the flood class. The reduction in the total number of observations is due to filtering pre-event data and excluding features that potentially contain information leaks. However, the balance between the classes is maintained.

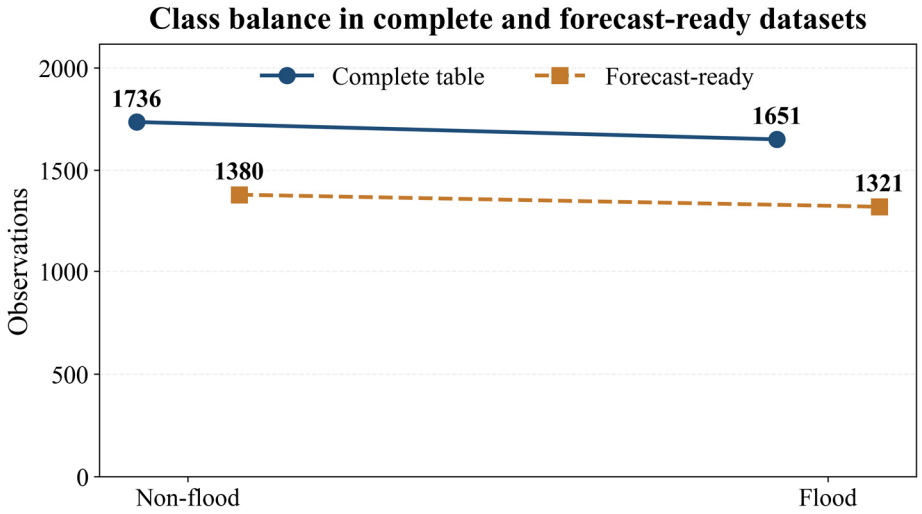

Figure 5. Class distributions in full and predictive-ready scientific datasets.

Thus, the presented methodology provides a holistic, reproducible approach to data preparation and analysis for early spatiotemporal flood hazard forecasting. The implemented leakage-safe pipeline ensures strict adherence to the forecasting framework by exclusively using pre-event features, event-correct sample division, and applying all preprocessing procedures (imputation, standardization, coding) only to the training data. An analysis of the dataset structure, including the distribution of observations by year, event dynamics, temporal phases, and class balance, demonstrates that the resulting sample is sufficiently complete, balanced, and representative for constructing machine learning models. Together, this provides the correct conditions for subsequent comparative analysis of algorithms and allows the obtained results to be interpreted as reflecting real patterns of flood hazard formation, rather than data artifacts or information leaks.

3.3. Architecture of the RS-ALOP-ML Hybrid Model

This subsection presents the architecture of the proposed hybrid RS-ALOP-ML model, designed for early flood hazard forecasting over multiple time horizons. The model is based on an ensemble of logistic experts that use the same feature space but differ in the level of parameter regularization. This approach allows combining the robustness of strongly regularized models and the flexibility of weakly regularized branches. The final solution is formed through adaptive probabilistic fusion of expert outputs, which provides a more stable risk assessment and preserves the interpretability of the model. Figure 6 shows the internal structure of the proposed hybrid RS-ALOP-ML model in a horizon-specific setting. The upper block, One horizon-specific RS-ALOP-ML instance, defines a separate instance of the model for a single time horizon $h \in \{1,\ 7,\ 14,\ 30\}$ and accepts as input a standardized feature matrix $X_{scaled} \in \mathbb{R}^{B\times 38}$, where B is the number of observations in the current batch or subsample, and 38 is the dimension of the feature space.

This notation means that each row of the matrix corresponds to one observation object, and each column corresponds to one predictive feature after leakage-safe preprocessing and scaling. The horizon parameter is included in the scheme not as an additional numerical feature, but as an index of a separate model instance. This means that each horizon has its own set of parameters and its own composition of probabilistic outputs. This division is reflected in the scheme itself by the notation (18):

$$X_{scaled}^{(h)} \in \mathbb{R}^{B\times 38}, h \in \{1,7,14,30\} \quad (18)$$

The next central block contains the ordering of the regularization coefficients (19):

$$\lambda_{strong} > \lambda_{ref} > \lambda_{weak} > 0 \tag{19}$$

This line defines not an abstract relation, but rather the precise logic behind the distinctions between the three branches of the model: all three experts have the same linear-logistic form, the same input feature space, and the same binary target, but differ in the degree of regularization compression of the parameters. Therefore, the distinction between the branches pertains not to the algorithm type, but to the coefficient estimation mode. In the notation of the diagram, the library parameter C is written as the reciprocal of λ (20):

$$C_j = \frac{1}{\lambda_j}, j \in \{\text{strong}, \text{ref}, \text{weak}\} \tag{20}$$

The left block of the Strong LR expert corresponds to the most strictly regularized logistic branch. For this branch, the linear predictor and probabilistic output are written as (21) and (22):

$$z_{strong} = X_{scaled}\beta_{strong} + b_{strong} \tag{21}$$

$$p_{strong} = \sigma(z_{strong}) = \frac{1}{1 + e^{-z_{strong}}}, p_{strong} \in [0,1]^B \tag{22}$$

Strong regularization here means that the coefficient vector β_{strong} is estimated under a more stringent norm constraint, so the weights of this branch are smoother and less sensitive to local feature fluctuations. The central block of Reference LR Expert defines an intermediate branch with a medium level of regularization (23) and (24):

$$z_{ref} = X_{scaled}\beta_{ref} + b_{ref} \tag{23}$$

$$p_{ref} = \sigma(z_{ref}), p_{ref} \in [0,1]^B \tag{24}$$

The right branch of Weak LR expert describes the weakest regularized logistic model, for which (25) and (26):

$$z_{weak} = X_{scaled}\beta_{weak} + b_{weak} \tag{25}$$

$$p_{weak} = \sigma(z_{weak}), p_{weak} \in [0,1]^B \tag{26}$$

All three blocks explicitly specify the same settings: class_weight = balanced and max_iter = 5000. This means that the difference between the branches is concentrated in the regularization parameter, rather than in the different class balancing schemes or different constraints on the optimization procedure. In analytical form, each of the three LR branches corresponds to a solution to the regularized logistic regression problem (27):

$$(\hat{\beta}_j, \hat{b}_j) = \arg\min_{\beta,b}\left[-\sum_{i=1}^{B}\left(y_i \log p_{ij} + (1-y_i)\log(1-p_{ij})\right) + \lambda_j \parallel \beta \parallel_2^2\right], j \in \{\text{strong}, \text{ref}, \text{weak}\} \tag{27}$$

where $p_{ij} = \sigma(x_i^\top \beta + b)$. The diagram thus shows that the three branches are three coordinated linear experts evaluating the same object using different coefficient smoothing modes. The intermediate Branch output blocks under each branch indicate that the output of each LR model is not represented by a binary decision, but by a probability vector of size B. These three probability vectors are fed to the Adaptive linear opinion pool block, where their union is defined by a convex linear combination (28):

$$p_{pool} = \omega_{strong}p_{strong} + \omega_{ref}p_{ref} + \omega_{weak}p_{weak}, \omega_j \geq 0, \sum_j \omega_j = 1 \tag{28}$$

Adaptive linear opinion pooling offers several theoretical advantages over traditional ensemble aggregation strategies. Unlike majority voting, it preserves the full probabilistic information generated by each expert, rather than reducing predictions to binary decisions. The weighted pooling scheme allows the three regularization-specific logistic experts to contribute unequally to the final probability estimate according to the predefined ALOP weighting configuration. Unlike stacking-based approaches, the proposed fusion mechanism does not require training an additional meta-learning algorithm, reducing the risk of overfitting while maintaining full probabilistic interpretability. Consequently, the proposed aggregation strategy achieves a favorable balance between robustness, transparency, and computational efficiency. The proposed RS-ALOP-ML framework should not be interpreted as a simple ensemble of independent logistic regression classifiers. Its methodological novelty lies in the coordinated integration of three complementary components: (i) leaky-robust model construction with a given forecast horizon, (ii) diversity generation using regularization-guided logistic regression experts instead of heterogeneous base learning models, and (iii) adaptive probabilistic fusion based on Adaptive Linear Opinion Pooling (ALOP), which preserves probabilistic calibration, model interpretability, and robustness across multiple forecast horizons. Unlike traditional ensemble methods that rely on heterogeneous classifiers, majority voting, or stacking-based meta-model learning, the proposed framework forms a unified probabilistic forecasting architecture specifically designed for multimodal environmental forecasting under strict evaluation conditions that ensure leaky-robustness. From this entry it follows that the aggregate output remains a probability and belongs to the same range (29):

$$p_{pool} \in [0,1]^B \tag{29}$$

The lower left block, "Probability output," represents this final vector of probability estimates without any additional transformation. The lower right block, "Binary early-warning output," specifies the threshold representation of the result (30):

$$\hat{y}_h = 1\left[p_{pool} \geq \tau_h^*\right] \in \{0,1\}^B \tag{30}$$

where τ_h^* is the decision threshold associated with a specific horizon h. The index h in this notation is crucial: it indicates that the final binary decision is associated not with a single global boundary, but with a horizon-specific model configuration. The arrows in the diagram represent a strictly sequential transition from a single input feature space to three parallel expert branches, then to an aggregated probabilistic representation, and finally to a binary early-warning output. The color-coded branches visually separate the three regularization modes, and the symmetrical arrangement of the blocks emphasizes that the branches are equivalent in architectural form and differ only in their parameterization. The entire diagram describes the proposed new hybrid model not as a set of unrelated algorithms, but as a single ensemble of logistic experts with controlled regularization and coordinated probabilistic unification. Thus, the RS-ALOP-ML model represents a coordinated hybrid architecture in which three logistic experts operate in parallel regularization modes to generate probabilistic flood risk assessments. Combining them via an adaptive linear opinion pool yields a final forecast that accounts for varying levels of robustness and feature sensitivity. Horizon-specific formulation further ensures model adaptation to each forecast time horizon. This makes RS-ALOP-ML not simply an ensemble of individual classifiers, but a holistic and interpretable flood early warning model. Unlike traditional ensemble methods that combine heterogeneous classifiers (e.g., Random Forest, boosting, or stacking), RS-ALOP-ML uses a coordinated ensemble of homogeneous logistic experts, differing only in their regularization strength. This design explicitly controls for model

diversity while maintaining a common probabilistic formulation and full interpretability of the coefficients.

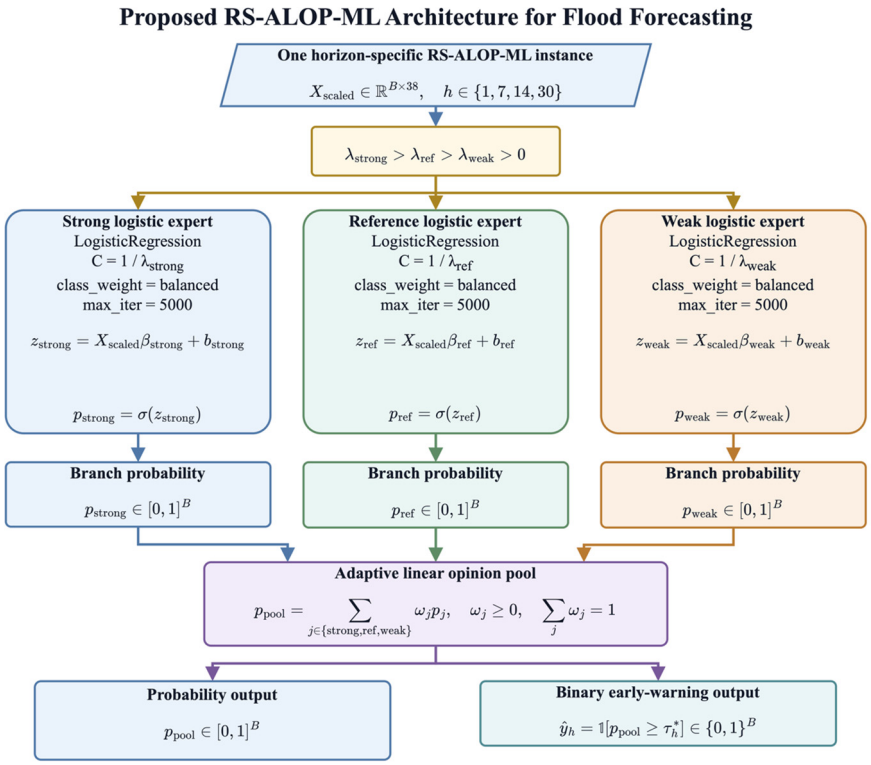

Figure 6. Architecture of the proposed hybrid Remote Sensing Adaptive Linear Opinion Pool Machine Learning (RS-ALOP-ML) model for early flood hazard forecasting.

Artificial intelligence (AI) tools were used solely to improve the linguistic quality, grammar, and readability of the manuscript. AI tools were not used for study design, data collection, data analysis, model development, results interpretation, or scientific decision making. All scientific content, methodology, experiments, validation, and conclusions were developed, reviewed, and approved by the authors.

4. Results

4.1. Spatio-Temporal Structure of the Forecast Sample and Event Representation

This section analyzes the spatiotemporal structure of the generated sample, including the distribution of observations across the territory, their association with individual events, and the specifics of generating forecast-oriented data subsets. Visualizing spatial coverage allows us to assess the representativeness of the sample, identify the distribution pattern of flood and control observations, and verify the validity of the event-oriented problem statement. Particular attention is paid to the consistency of the geographic distribution of data across different forecast time horizons, which is critical for interpreting the modeling results. Figure 7 shows that the documented floods are geographically distributed across several major river basins and climatic regions of Kazakhstan, rather than concentrated in a single area. This spatial diversity increases the geographic representativeness of the dataset and allows the proposed model to study flood-related patterns in a variety of environmental conditions.

This visualization is meaningful for several reasons. First, it confirms that the data are not limited to a single local area but span several hydrologically distinct zones of Kazakhstan, including the northern, western, central, eastern, and southern regions. Second, the map shows that positive and negative observations coexist in close spatial proximity,

rather than being separated by the "one region, one class" principle. This reduces the risk that the model will rely solely on coarse geographic localization instead of analyzing meaningful hydrometeorological and geospatial features. Third, plotting event centroids and local events demonstrates the association of table rows with specific flood episodes, which is particularly important for event-based problem formulation and subsequent grouping by event_id. The map also allows us to conclude the practical logic behind the sample construction: flood rows were formed around actual spatiotemporal hazard centers, and control observations were used to create a comparable background, necessary for accurate binary forecasting. The presence of locally curated events complements the main event contour and expands the representativeness of the sample at later observation intervals. Taken together, the presented geography confirms that the studied dataset has an event-based, spatially consistent, and applied structure suitable for spatiotemporal modeling of early flood risk.

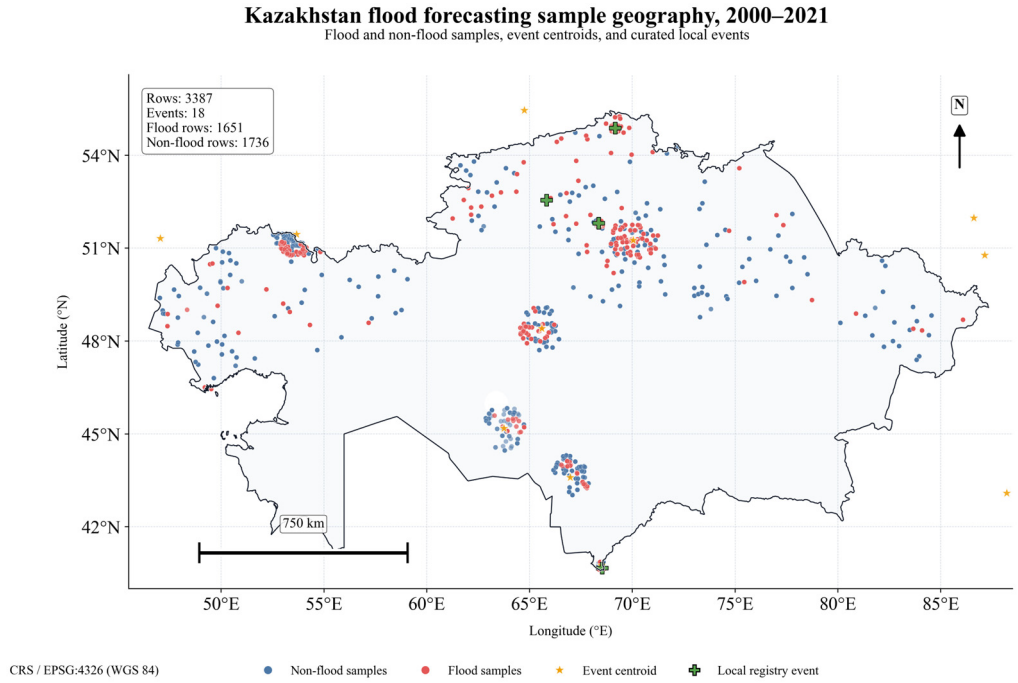

Figure 7. Spatial distribution of floods. Red circles represent floods, blue polygons represent permanent water bodies, and the background shows elevation.

Figure 8 demonstrates that the four forecasting horizons (T + 1, T + 7, T + 14, and T + 30) preserve a consistent spatial distribution of flood and non-flood observations across the principal flood-prone regions of Kazakhstan. Although the number of observations varies slightly between horizons (679, 655, 653, and 714, respectively), the geographical coverage and class balance remain comparable. This consistency indicates that differences in predictive performance across forecasting horizons are primarily associated with the availability of pre-event information rather than changes in the spatial composition of the dataset, thereby supporting a fair comparison of model performance under leakage-safe evaluation conditions.

Figure 9 shows the spatial structure of the representative flood events included in the dataset. Observations of floods and non-flood events are distributed within the same regions, providing geographically consistent positive and control samples for model development. The selected events exhibit varying spatial extent and density of observations, reflecting the natural variability of flood occurrence in Kazakhstan. This event-based sampling strategy allows the model to learn discriminatory patterns associated with floods

while maintaining local geographic context and ensuring safe separation of flood and non-flood event observations, taking into account leakage risk.

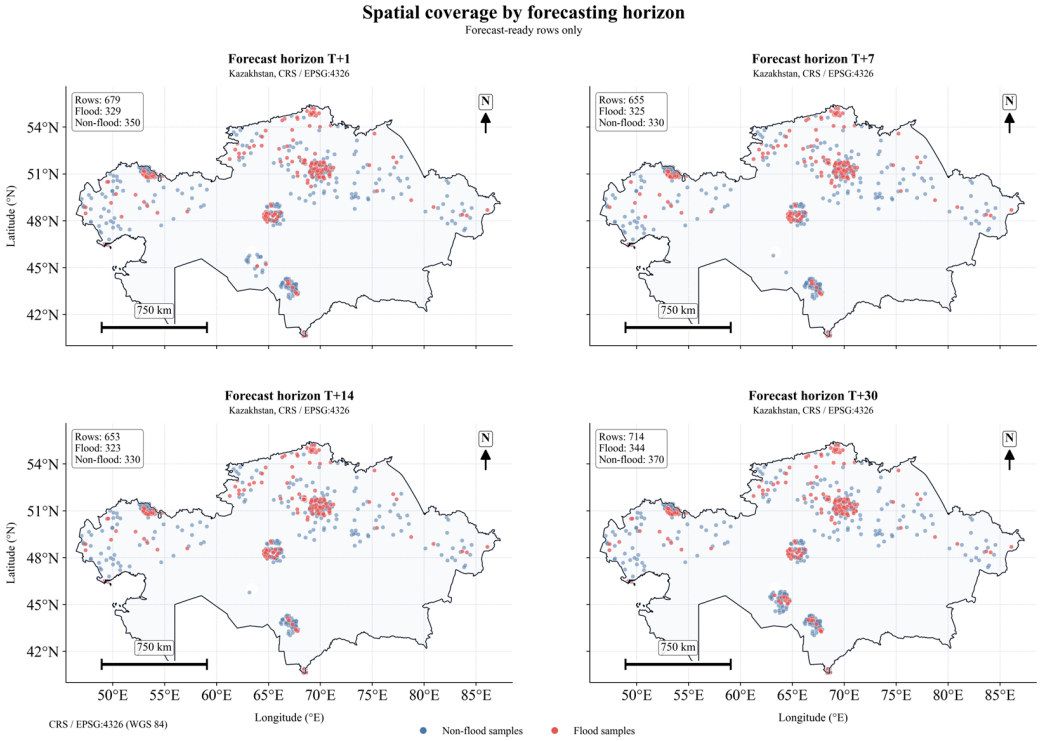

Figure 8. Spatial coverage of the forecast sample for early flood warning horizons.

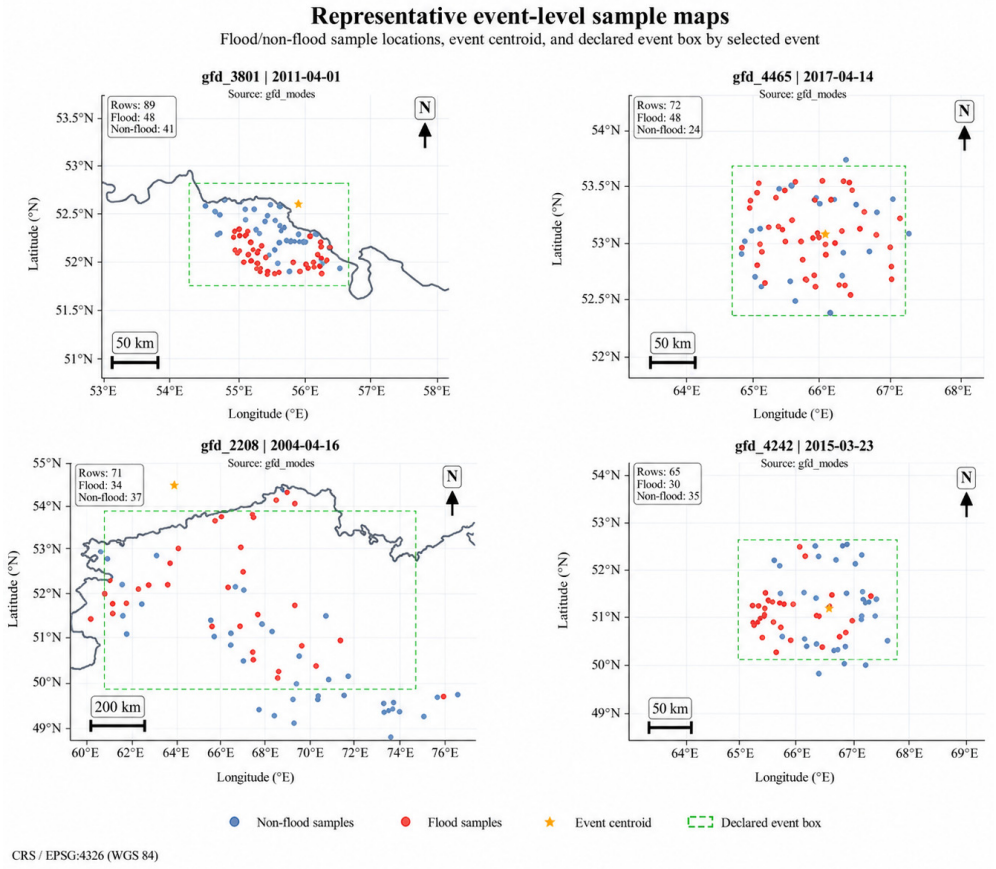

Figure 9. Red dots represent flood observations, blue dots represent non-flood observations, a yellow star indicates the event center, and a green dotted rectangle represents the event's original boundary. Geographic coordinates are given in WGS 84 (EPSG:4326).

Figure 10 shows the local event registry entries plotted on a map of Kazakhstan in the CRS/EPSG:4326 (WGS 84) geographic coordinate system. Green crosses denote locally recorded flood events. Each mark corresponds to a single event, and the captions next to it contain its identifier, including the year and a short location name. Four events are represented on the map: 2019_nko_ishim, 2021_karamyrza, 2021_atbasar_zhabay, and 2020_maktaaral. The green dotted rectangles correspond to the declared event box, that is, the declared spatial boundaries of the event. These contours show the boundaries of the area associated with a specific local registry entry. If the box is very small relative to the scale of the entire map, it is visually perceived as a compact dotted outline around the event point.

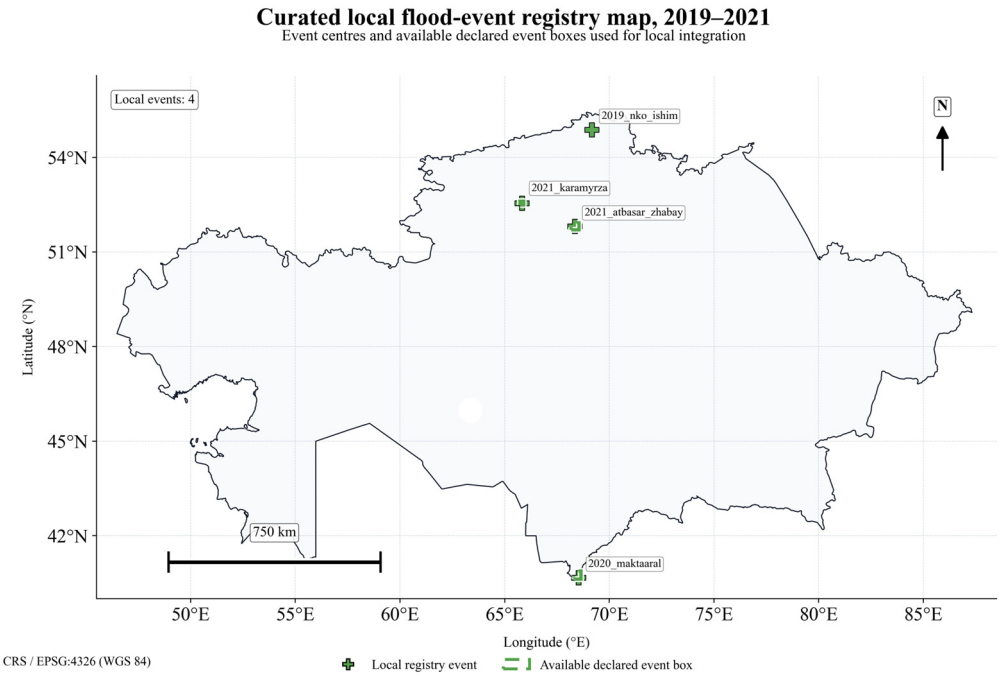

Figure 10. The map shows the state border of Kazakhstan, geographic coordinates (WGS 84, EPSG:4326), an arrow pointing north, a 750 km scale bar, and the locations of four locally recorded floods.

Spatial analysis confirms that the created dataset has a consistent geographic structure and event-driven organization. Floods are distributed across the main flood-prone regions of Kazakhstan, with observations of floods and unrelated events coexisting in the same local areas, reducing the risk of geographic sampling bias. Furthermore, consistent spatial coverage maintained across all forecast horizons ensures fair comparison of models and indicates that differences in forecast performance primarily reflect changes in forecasting complexity rather than variations in the geographic composition of the dataset. Overall, these results confirm the suitability of the proposed dataset for developing robust and generalizable early flood hazard forecasting models.

4.2. Analysis of Feature Space and Dependencies

This subsection conducts a statistical analysis of the resulting feature space to identify the distributional characteristics of the variables, their interrelations, and the degree of information they provide relative to the target variable. This analysis is an important stage of data preparation, as it allows for assessing feature heterogeneity, identifying potential multicollinearity, and determining which variables contribute most to flood risk. Descriptive statistics and linear and nonlinear correlation analysis methods are used for this purpose, providing a comprehensive assessment of the data structure. Table 1 shows that the features exhibit significantly different scales and distribution patterns. The widest dispersion is

observed for the relief and hydrometeorological variables: for elevation, the interquartile range is 203, and for precip_sum_30d, it is 21.5252, indicating pronounced spatial and temporal heterogeneity of conditions. All hydrometeorological variables are presented in their original physical units (e.g., precipitation in millimeters (mm), temperature and dew point in degrees Celsius (°C), snow depth in meters (m), and volumetric soil moisture in m^3/m^3), while spectral indices are dimensionless. The water variables water_occurrence, water_recurrence, water_seasonality, and binary indicators such as max_water_extent or permanent_water_mask have medians close to zero, indicating a predominance of non-flooded or irregularly flooded states with a limited number of water-saturated locations.

Table 1. Descriptive statistics of the main features of the final data set.

Feature	Mean	Std	Min	q1	Median	q3	Max	Interquartile Range (q3 − q1)
elevation	228.0796	166.6776	−31.0000	104.0000	209.0000	307.0000	1709.0000	203.0000
aspect	153.0459	113.8995	0.0000	41.0000	158.0000	258.0000	356.0000	217.0000
day_of_year	92.8327	43.0081	16.0000	76.0000	92.0000	105.0000	365.0000	29.0000
water_recurrence	21.0207	32.6892	0.0000	0.0000	0.0000	45.0000	100.0000	45.0000
precip_sum_30d	29.1750	18.8579	4.0125	15.4822	24.4814	37.0074	144.3298	21.5252
water_occurrence	4.6246	13.6648	0.0000	0.0000	0.0000	2.0000	96.0000	2.0000
precip_sum_14d	13.9304	12.5824	0.2809	4.3918	11.3216	19.2626	86.6021	14.8708
temp_mean_3d	0.1832	8.5424	−20.0384	−4.1573	−0.9758	3.4879	25.7113	7.6452
valid_obs_count	20.3669	8.4426	4.0000	14.0000	21.0000	27.0000	42.0000	13.0000
temp_mean_7d	−0.4537	8.2949	−22.1508	−4.6191	−1.2359	3.4622	25.9895	8.0813
precip_sum_7d	6.2925	7.7521	0.0053	0.7889	3.7884	10.0807	61.8675	9.2918
pre_event_wetness_proxy	37.3913	6.8778	14.5772	32.3584	37.3450	41.3275	59.8669	8.9691
dewpoint_mean_7d	−5.2052	6.8495	−24.8522	−8.6194	−5.6192	−2.1374	18.9170	6.4820
hydro_meteo_risk_proxy	11.8113	3.3954	3.0571	9.4170	11.4298	13.9694	35.9978	4.5524
precip_sum_3d	2.3038	3.2298	0.0009	0.0613	1.1383	3.4576	31.9626	3.3963
water_seasonality	0.6564	1.7682	0.0000	0.0000	0.0000	1.0000	12.0000	1.0000
month	3.5372	1.3938	1.0000	3.0000	4.0000	4.0000	12.0000	1.0000
pre_event_precipitation_intensity	0.7679	1.0766	0.0003	0.0204	0.3794	1.1525	10.6542	1.1321
soil_texture_class	6.9911	0.8027	4.0000	7.0000	7.0000	7.0000	12.0000	0.0000
slope	0.2643	0.6404	0.0000	0.0000	0.0000	0.0000	5.0000	0.0000
MNDWI	0.4888	0.5008	−0.628	0.3012	0.7635	0.8123	0.9162	0.5111
NDMI	0.5511	0.3795	−0.340	0.3908	0.7547	0.8011	0.9082	0.4102
max_water_extent	0.0881	0.2835	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
BSI	−0.1572	0.1908	−0.310	−0.2730	−0.2630	−0.1389	0.3380	0.1341
reference_month_water	0.0307	0.1726	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
EVI	−0.1075	0.1678	−0.598	−0.2365	−0.1163	0.0554	0.6369	0.2919
snow_depth_mean_7d	0.1514	0.1442	−0.000	0.0083	0.1284	0.2673	1.2167	0.2590
NDWI	−0.0542	0.1407	−0.720	−0.0880	0.0192	0.0355	0.1409	0.0
permanent_water_mask	0.0178	0.1321	0.0000	0.0000	0.0000	0.0000	1.0000	0.0
NDVI	0.0313	0.1052	−0.110	−0.0319	−0.0189	0.0455	0.8336	0.0
previous_month_water	0.0056	0.0743	0.0000	0.0000	0.0000	0.0000	1.0000	0.0
soil_moisture_mean_7d	0.3586	0.0694	0.1223	0.3048	0.3530	0.4021	0.5913	0.0
soil_moisture_mean_14d	0.3542	0.0678	0.1419	0.3030	0.3464	0.3954	0.5759	0.0

The spectral indices MNDWI and NDMI exhibit significant variability, while NDVI, NDWI, and BSI are concentrated in narrower ranges. For soil_texture_class and slope, the

interquartile range is zero, indicating a strong concentration of values in one dominant interval. Crucially, for all presented features, missing_share = 0.0, meaning the predictor working space is free of gaps and complete at the modeling stage. These characteristics together confirm that the final dataset combines diverse features with different variance, skewness, and physical range and, therefore, requires careful scale normalization and interpretation of variable contributions during subsequent model comparisons. Figure 11 shows a heat map of pairwise Pearson correlation coefficients between the selected forecast-ready predictors.

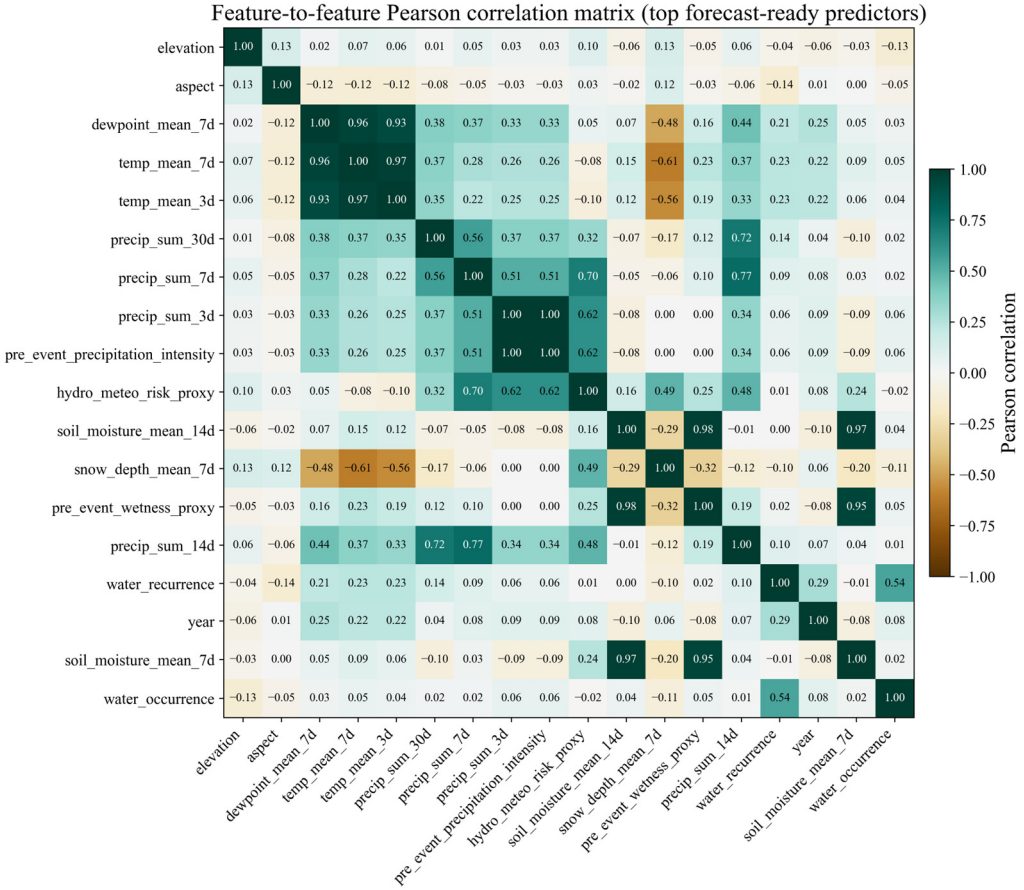

Figure 11. Pearson matrix of pairwise correlations between the most informative predictive features.

The same features are listed on both axes. Hence, each cell reflects the relationship between a pair of variables, and the main diagonal is filled with values of 1.00, corresponding to a perfect correlation between the feature and itself. The color scale on the right defines the coefficient range from −1.00 to 1.00: dark green shades indicate a strong positive relationship, light neutral tones indicate a weak or near-zero relationship, and brown-orange shades indicate a negative relationship. The numerical value of the coefficient is plotted within each cell, allowing for the precise strength and sign of the relationship to be read.

The strongest positive relationships are observed in the temperature block: dewpoint_mean_7d, temp_mean_7d, and temp_mean_3d form a tight cluster with coefficients of 0.93–0.97. In the precipitation block, pronounced positive dependencies are observed between precip_sum_30d and precip_sum_14d (0.72), precip_sum_14d and precip_sum_7d (0.77), and precip_sum_3d and pre_event_precipitation_intensity (1.00), indicating near-perfect linear coincidence. The hydro_meteo_risk_proxy feature is positively associated with short-term precipitation variables, including precip_sum_7d (0.70) and precip_sum_3d/pre_event_precipitation_intensity (0.62 each).

Soil moisture characteristics and moisture proxies form a separate cluster: soil_moisture_mean_14d is closely related to pre_event_wetness_proxy (0.98) and soil_moisture_mean_7d (0.97), and pre_event_wetness_proxy is closely related to soil_moisture_mean_7d (0.95). The water features water_recurrence and water_occurrence have a moderately positive relationship of 0.54. Among the negative dependencies, the most notable correlations are with snow_depth_mean_7d: −0.61 with temp_mean_7d, −0.56 with temp_mean_3d, and −0.48 with dewpoint_mean_7d. The relief features' elevation and aspect generally demonstrate weak relationships with most other variables, as reflected by coefficients close to zero. Figure 12 shows a heatmap of the point-by-point correlation between each predictor feature and the binary target variable. The vertical axis lists the predictors, and a single column displays the R value for the association with the flood/non-flood target class. The color scale on the right indicates the sign and magnitude of the association: red shades correspond to a positive association with the flood class, blue shades to a negative association, and light neutral shades to values close to zero. The exact numerical value of the coefficient is given within each cell.

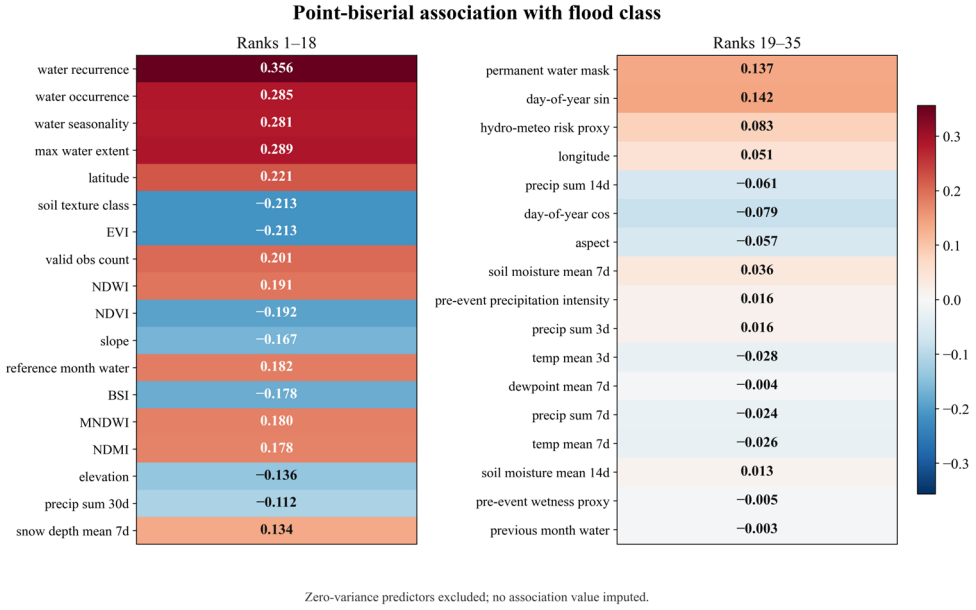

Figure 12. Linear relationship matrix of predictive features with a binary flood target variable.

The highest positive values are observed for water_recurrence (0.356), year (0.344), max_water_extent (0.289), water_occurrence (0.285), and water_seasonality (0.281). This means that these features demonstrate the strongest linear positive association with the flood label. In the same group, but with a more moderate correlation strength, are valid_obs_count (0.201), NDWI (0.191), reference_month_water (0.182), MNDWI (0.180), NDMI (0.178), permanent_water_mask (0.137), and snow_depth_mean_7d (0.134). The hydro_meteo_risk_proxy feature has a weak positive correlation (0.083), and soil_moisture_mean_7d, pre_event_precipitation_intensity, precip_sum_3d, and soil_moisture_mean_14d are very close to zero.

Slightly negative values are present for previous_month_water (−0.003), dewpoint_mean_7d (−0.004), pre_event_wetness_proxy (−0.005), month (−0.016), day_of_year (−0.022), precip_sum_7d (−0.024), temp_mean_7d (−0.026), and temp_mean_3d (−0.028). A more pronounced negative association is observed for aspect (−0.057), precip_sum_14d (−0.061), precip_sum_30d (−0.112), elevation (−0.136), slope (−0.167), BSI (−0.178), NDVI (−0.192), EVI (−0.213), and soil_texture_class (−0.213). The most saturated red and blue cells visually highlight the features with the strongest absolute linear relationships.

Figure 13 shows a heat map of the Spearman rank correlation coefficients between each predictor feature and the binary target variable. The vertical axis lists the predictors, and a single column displays the ρ value for the relationship with the flood/non-flood target class. The color scale on the right indicates the sign and magnitude of the association: green shades correspond to a positive monotonic relationship, purple shades to a negative one, and light intermediate shades indicate a weak or near-zero relationship. The exact numerical value of the coefficient is indicated within each cell.

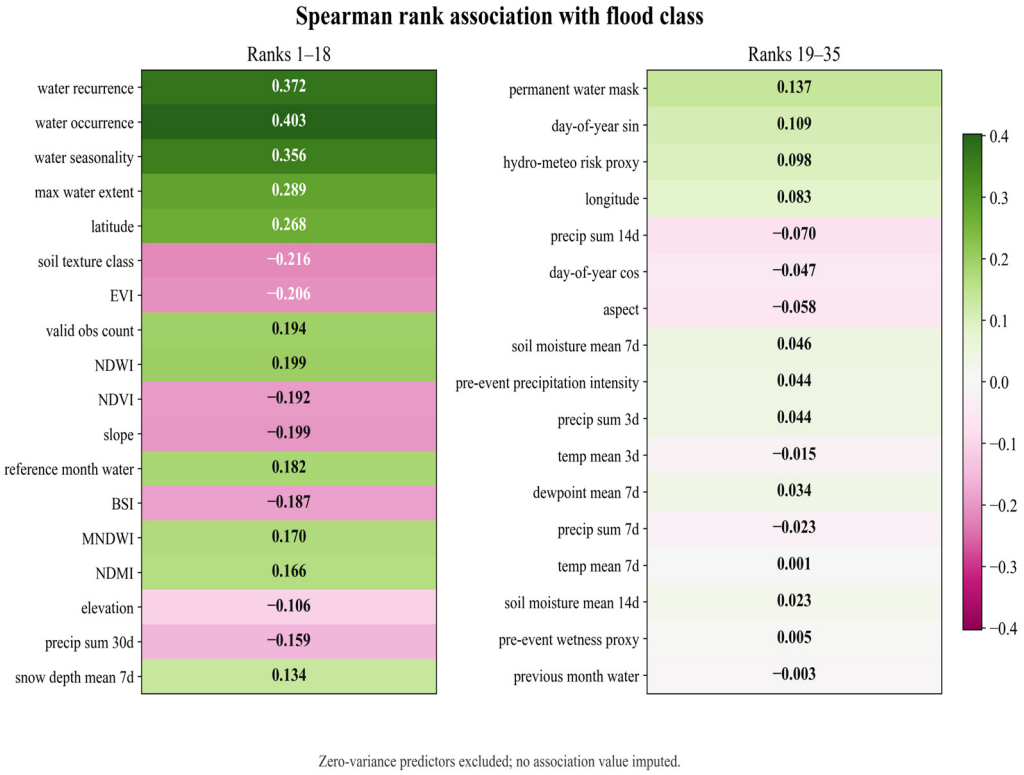

Figure 13. The matrix of monotonic relationships of forecast features with the binary target variable of flood according to Spearman's coefficient.

The most pronounced positive values are observed for water_occurrence (0.403), water_recurrence (0.372), water_seasonality (0.356), year (0.340), and max_water_extent (0.289). These features are located at the top of the matrix and are highlighted in the most saturated green tones. They are followed by NDWI (0.199), valid_obs_count (0.194), reference_month_water (0.182), MNDWI (0.170), NDMI (0.166), permanent_water_mask (0.137), and snow_depth_mean_7d (0.134). Weaker but positive associations are shown for hydro_meteo_risk_proxy (0.098), soil_moisture_mean_7d (0.046), pre_event_precipitation_intensity (0.044), precip_sum_3d (0.044), dewpoint_mean_7d (0.034), month (0.028), soil_moisture_mean_14d (0.023), day_of_year (0.018), pre_event_wetness_proxy (0.005), and temp_mean_7d (0.001).

The negative portion of the matrix begins with a near-zero value for previous_month_water (−0.003), followed by temp_mean_3d (−0.015), precip_sum_7d (−0.023), aspect (−0.058), precip_sum_14d (−0.070), elevation (−0.106), and precip_sum_30d (−0.159). The most pronounced negative coefficients are for BSI (−0.187), NDVI (−0.192), slope (−0.199), EVI (−0.206), and soil_texture_class (−0.216), reflected by the most saturated purple hues at the bottom of the matrix.

Figure 14 shows a heatmap of the distance correlation values between each predictor feature and the binary flood target variable. The vertical axis lists the predictors, and the single column indicates the magnitude of the nonlinear association. The color scale on the right shows the strength of the relationship: lighter yellow-green shades indicate higher

values, while darker blue-violet shades indicate lower values. The exact numerical value of the coefficient is given within each cell.

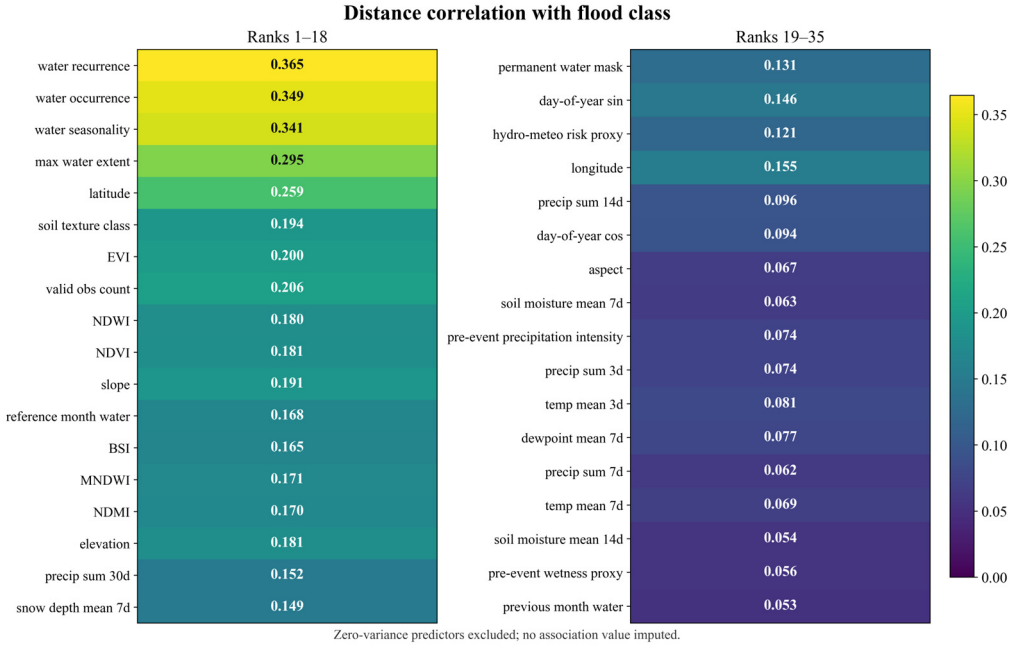

Figure 14. The matrix of nonlinear relationships between predictive features and a binary target variable using distance correlation.

The highest values are observed for water_recurrence (0.365), water_occurrence (0.349), water_seasonality (0.341), year (0.338), and max_water_extent (0.295). These features form the top of the matrix and are the most closely related to the target variable. The next level consists of valid_obs_count (0.206), EVI (0.200), soil_texture_class (0.194), slope (0.191), NDVI (0.181), elevation (0.181), NDWI (0.180), MNDWI (0.171), NDMI (0.170), reference_month_water (0.168), and BSI (0.165). For these variables, the relationship is moderate. Average and weaker values are shown for precip_sum_30d (0.152), snow_depth_mean_7d (0.149), permanent_water_mask (0.131), hydro_meteo_risk_proxy (0.121), precip_sum_14d (0.096), day_of_year (0.092), month (0.087), temp_mean_3d (0.081), dewpoint_mean_7d (0.077), precip_sum_3d (0.074), pre_event_precipitation_intensity (0.074), temp_mean_7d (0.069), aspect (0.067), soil_moisture_mean_7d (0.063), precip_sum_7d (0.062), pre_event_wetness_proxy (0.056), soil_moisture_mean_14d (0.054) and previous_month_water (0.053). The lower part of the matrix contains features with the least nonlinear association, while the upper part reflects variables with the most pronounced dependence on the target class.

Thus, the analysis shows that the resulting feature space is characterized by highly heterogeneous distributions and by both strong and weak dependencies between variables. Features related to long-term water dynamics and the area's hydrological predisposition are the most informative. At the same time, hydrometeorological and spectral indicators exhibit more complex, often nonlinear relationships with the target variable. Groups of correlated features were also identified, confirming the need to use regularization and ensemble approaches when constructing models. Taken together, the results substantiate the chosen modeling methodology and confirm that flood hazard forecasting requires consideration of both linear and nonlinear patterns in the data. To assess the statistical robustness of the proposed framework, bootstrap resampling was additionally applied to the main evaluation metrics. The corresponding mean values and 95% confidence intervals

for ROC-AUC, PR-AUC, F1-score, balanced accuracy, MCC, and Brier score are reported in Table S3.

4.3. Modeling Results and Interpretation of Features

This section presents the modeling results and their interpretation for the problem of early flood hazard forecasting. The analysis includes a comparative assessment of the quality of various machine and deep learning models, as well as a study of the contribution of individual features to forecasting. Classification metrics, ablation experiments, and interpretation methods, including SHAP analysis, are used for this purpose. This integrated approach not only identifies the most effective model but also explains the factors underlying the obtained results. To interpret the model results and assess the contribution of individual features to forecasting, an analysis based on the SHAP (SHapley Additive exPlanations) method was conducted. This approach allows a quantitative assessment of each feature's influence on the final probability of classifying an observation as a flood hazard and identifies both its direction and magnitude. Figure 15 shows SHAP beeswarm diagrams for two variants of the training set: the forecast-ready subset and the final dataset. Each point on the graph corresponds to a single observation, and its position along the x-axis reflects the feature's contribution to flood probability. The color scale indicates the feature's value, from low (blue) to high (red). The analysis shows that the most significant contributions to the forecast come from features related to long-term water dynamics and the area's hydrological predisposition, such as water_occurrence, water_recurrence, and water_seasonality. Topographic characteristics, particularly elevation, also play a significant role, while aggregated precipitation indicators, such as precip_sum_30d, stand out among hydrometeorological factors. A comparison of the two diagrams reveals that the feature-importance structure is generally preserved; however, the forecast-ready dataset shows a more pronounced concentration of contributions around key variables, due to the exclusion of post-event information and the shift to a strictly pre-event problem statement. In the full dataset, the contribution of individual features may be more smoothly distributed due to additional observations.

Figure 16 presents the comparative performance of all evaluated models on the time-lag set using ROC-AUC, PR-AUC, and Brier score. RS-ALOP-ML achieved the highest ROC-AUC (0.780), slightly outperforming logistic regression (0.779) and Process Wide&Deep (0.777), while RS-ALOP-ML and logistic regression obtained the highest PR-AUC (0.930). RS-ALOP-ML also obtained the lowest Brier score (0.212), indicating the most accurate probabilistic predictions among the evaluated models. In contrast, the deep tabular models FT-Transformer and TabNet showed significantly lower discrimination performance: ROC-AUC values were 0.658 and 0.563, and PR-AUC values were 0.849 and 0.808, respectively. Their higher Brier scores (0.316 and 0.290) further indicate less reliable probability estimates under temporal stagnation conditions. Such probabilistic forecasts are particularly suitable for operational decision support, where emergency response measures depend on estimated risk levels rather than deterministic class assignments. Similar observations regarding the practical role of predictive analytics in decision support have recently been noted in other machine learning application areas [40]. To determine whether the observed performance differences were robust to sample variability, an additional bootstrap comparison was conducted between RS-ALOP-ML and the closest competing baseline models. The mean performance differences and their corresponding 95% confidence intervals are presented in Table S4.

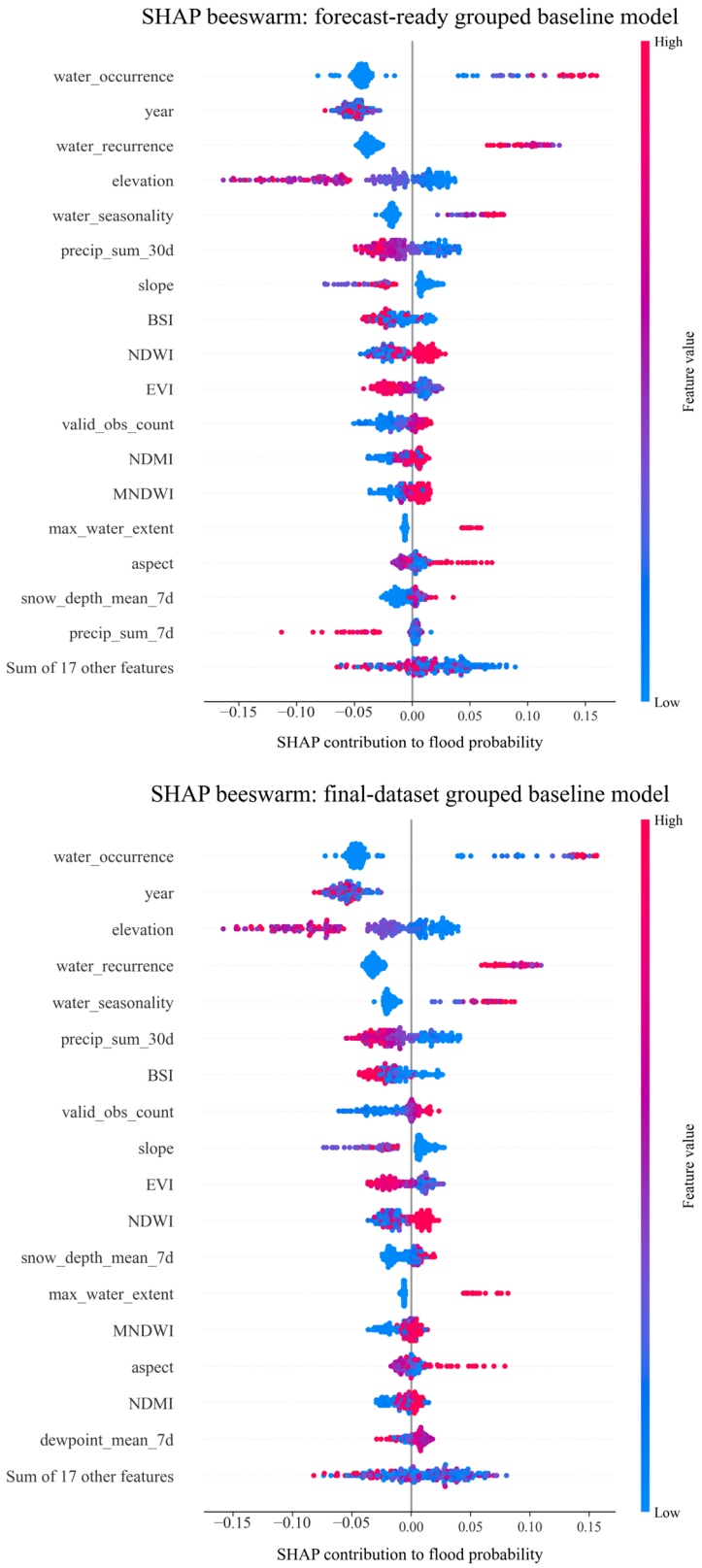

Figure 15. SHAP analysis of the contribution of features to flood hazard forecasting for full and forecast-ready datasets.

Figure 17 presents the results of the ablation analysis of the components of the proposed RS-ALOP-ML model using the ROC-AUC metric for different forecasting horizons (T + 1, T + 7, T + 14, and T + 30 days). The analysis includes individual model branches (Strong LR, Reference LR, and Weak LR), their equal-weight pooling (Equal-weight pool-

ing), and the final hybrid RS-ALOP-ML model with adaptive probability pooling. The results show that individual logistic branches demonstrate similar ROC-AUC values across all forecasting horizons, ranging from approximately 0.729 to 0.772. This confirms that each branch is a robust base classifier, but their individual performance is limited. Equal-weight pooling yields a slight improvement in quality over individual branches, suggesting complementary information across the models. However, the best results are achieved with the proposed adaptive fusion method (RS-ALOP-ML), which consistently yields higher ROC-AUC values across all forecast horizons. Analysis by time horizon shows that forecast quality gradually improves with increasing horizon: from values of approximately 0.73–0.74 for T + 1 to approximately 0.76–0.77 for T + 30. This may indicate that earlier pre-event conditions contain more robust signals associated with flood hazard formation.

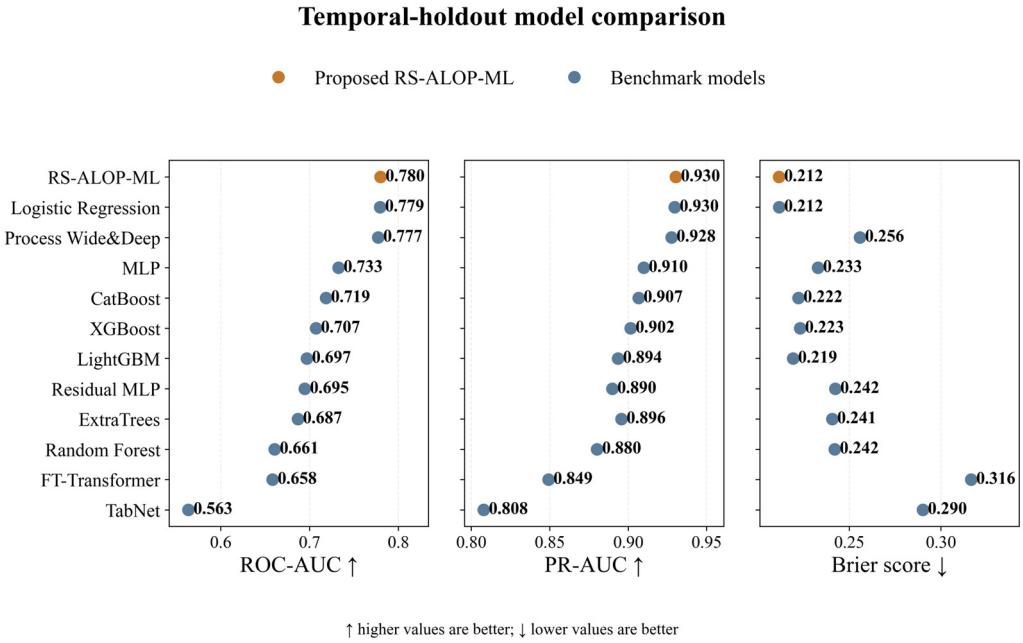

Figure 16. Comparison of models on an independent time-varying test set.

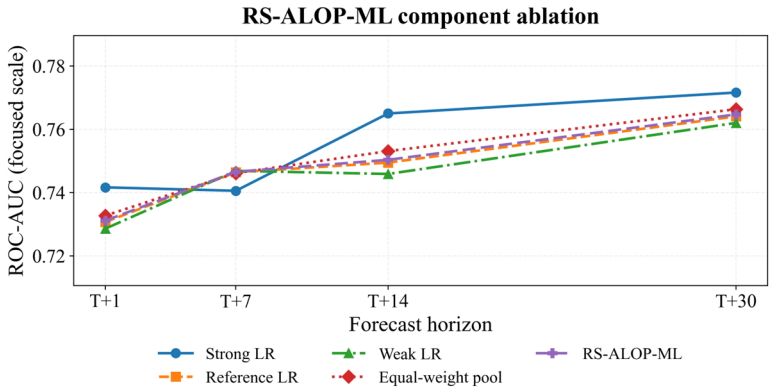

Figure 17. Ablative analysis of the components of the proposed RS-ALOP-ML model using the ROC-AUC metric at different forecasting horizons.

This ablation experiment was specifically designed to distinguish the contribution of the proposed architecture from that of its previously established individual components. Since logistic regression, regularization, and probability averaging are not new in themselves, the relevant question is whether their coordinated configuration provides an advantage over their constituent alternatives. Therefore, the comparison moves from individual regularization-specific experts to fixed-weight fusion and, finally, to the pro-

posed adaptive opinion fusion configuration. The consistently higher ROC-AUC obtained by RS-ALOP-ML across four forecast horizons indicates that the observed improvement cannot be attributed solely to the use of logistic regression or simple ensembling; rather, it is due to the adaptive fusion of intentionally diversified regularization-specific experts within the forecasting framework specific to each horizon.

Figure 18 shows how excluding individual feature groups affects forecast quality, measured by ROC-AUC, at the T + 1, T + 7, T + 14, and T + 30 horizons. The full set of features yields stable ROC-AUC values of 0.731–0.765. The most significant decrease in quality is observed when excluding hydrometeorological features: ROC-AUC drops to 0.491–0.607. This confirms that precipitation, temperature, soil moisture, and related indicators are key predictors of flood hazard. Removing relief, geographic, spectral, and temporal/qualitative features results in a smaller decrease or similar metric values. Consequently, these groups of features complement the model, but the hydrometeorological block makes the main contribution to the forecast.

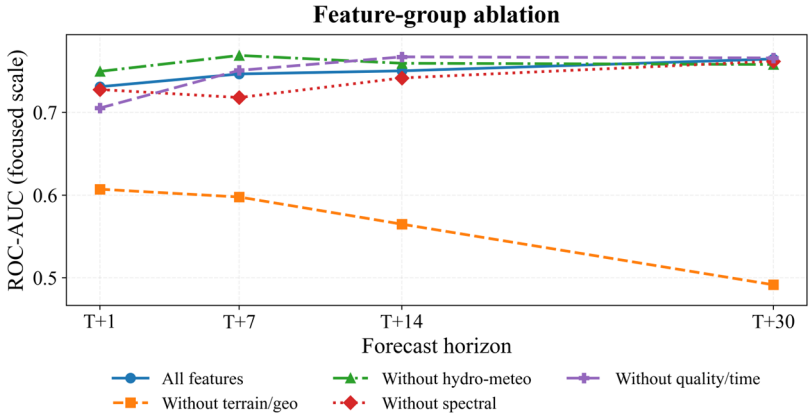

Figure 18. Ablative analysis of feature groups for the proposed model using the ROC-AUC metric at different forecasting horizons.

Figure 19 shows a horizontal bar chart of the mean absolute SHAP values for the most significant features of the RS-ALOP-ML model. The ordinate axis lists the features, and the abscissa axis shows the Mean |SHAP value|, which reflects the average contribution of a feature to the prediction, regardless of the direction of influence. The "elevation" feature makes the largest contribution, with a value of 0.089, underscoring its key role in flood hazard forecasting. These are followed by temp_mean_3d (0.069), longitude (0.062), NDMI (0.062), water_occurrence (0.062), soil_moisture_mean_7d (0.061), snow_depth_mean_7d (0.060), temp_mean_7d (0.059), water_recurrence (0.057), dewpoint_mean_7d (0.050), EVI (0.048), and precip_sum_7d (0.048). The ranking structure shows that the model does not rely on a single dominant source of information but instead uses a combination of several groups of factors: relief, geographic, hydrological, hydrometeorological, and spectral. Close SHAP values for longitude, NDMI, water_occurrence, soil_moisture_mean_7d, snow_depth_mean_7d, and temp_mean_7d indicate the formation of a stable layer of complementary traits after the leading factor, elevation.

The high performance of RS-ALOP-ML is explained by its reliance on a comprehensive description of the flood process, including topography, spatial position, long-term water dynamics, surface moisture conditions, snow cover, and temperature conditions. This distribution of contributions confirms the model's physical interpretability and its ability to generate a more robust flood hazard forecast. Thus, the obtained results demonstrate that the proposed RS-ALOP-ML model provides a consistent improvement in forecast quality compared to individual models and their simple ensembles. Ablation experiments confirm the contribution of both the model's architectural components and various groups

of features, with hydrometeorological and water characteristics being the most significant. SHAP analysis further demonstrates that the model relies on physically interpretable and complementary factors, including topography, water dynamics, and pre-event conditions. Taken together, this confirms that the achieved accuracy is not due to random dependencies, but reflects the real patterns of flood hazard formation, which makes the proposed approach reliable and applicable in early warning tasks.

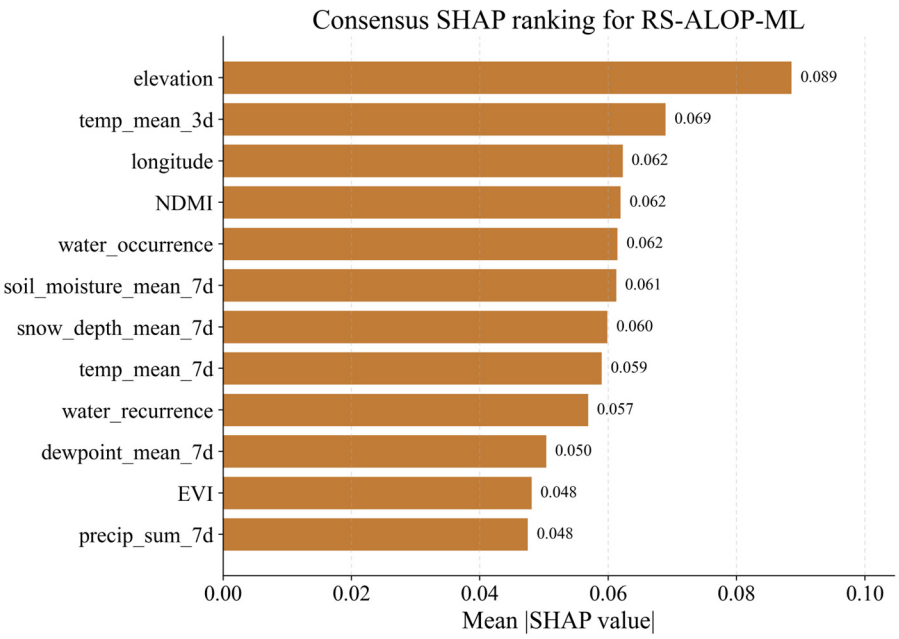

Figure 19. Consensus SHAP ranking of the most significant features in the proposed RS-ALOP-ML model.

Figure 20 compares the contributions of static and dynamic predictor groups with a full set of multimodal features across four forecast horizons. In grouped out-of-fold (OOF) event validation, the full model consistently outperformed configurations with only static and only dynamic predictors, while the dynamic predictors demonstrated a progressive decrease in ROC-AUC with increasing forecast horizon. In temporal evaluation with lagged sampling, the full and only static predictor configurations demonstrated relatively stable performance across all horizons, while the model with only dynamic predictors showed greater variability and remained consistently inferior to the full model.

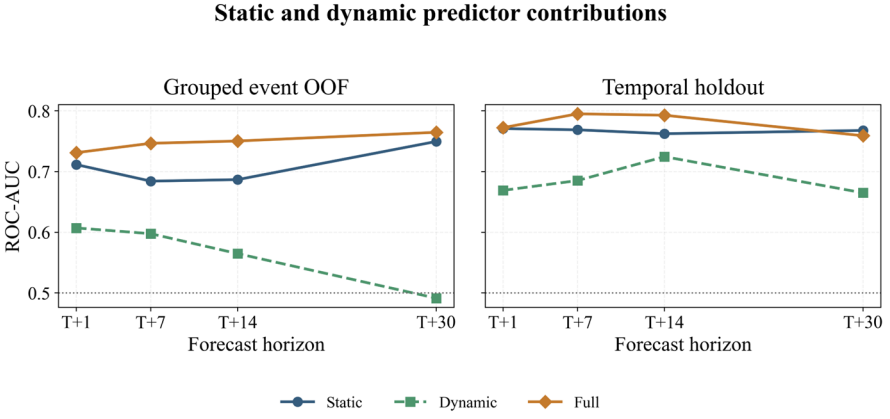

Figure 20. Contribution of static and dynamic predictors.

The results indicate that static geospatial and physiographic characteristics provide a stable baseline forecast model, while dynamic hydrometeorological variables provide additional event-specific information. Their combined use in a full feature configuration generally provides the most robust performance across both validation and forecast horizons.

5. Discussion

The main contribution of this study is not the introduction of a new hydrological forecasting algorithm, but the development of a leakage-safe probabilistic model for operational flood hazard assessment that combines multimodal geospatial information, event-oriented data organization, and adaptive probabilistic fusion. Instead of explicitly modeling the temporal evolution of flood processes, the proposed model estimates the probability of flood hazard based on available data on pre-event environmental conditions over different forecast horizons. This formulation is particularly suitable for operational early warning, where reliable probabilistic risk assessment and robust model generalization are often more important than deterministic forecasting of individual hydrological variables.

The results demonstrate that early flood hazard forecasting using multimodal spatiotemporal data requires not only high model accuracy but also a well-formulated problem. Unlike many existing studies that allow the use of post-event information, this study implements a strict leakage-safe approach, using exclusively pre-event features. This fundamentally affects the interpretation of the results, as the achieved metric values reflect the model's true predictive ability, not the effect of data leakage. A comparative analysis of the models showed that classical machine learning methods maintain high efficiency when working with tabular spatiotemporal data and, in some cases, outperform modern deep architectures. This is consistent with the results of recent benchmarking studies, which note the lack of a universal advantage of deep learning on medium-sized structured data. In this problem, tree-based methods and logistic regression yield stable results, whereas deep models exhibit greater variability and are more sensitive to hyperparameter tuning. This indicates that robust, interpretable models should be prioritized in data-constrained environmental modeling.

The proposed RS-ALOP-ML model demonstrates systematic performance improvements compared to both individual models and simple ensembles. Ablation analysis confirms that the key factor lies not simply in the use of an ensemble, but in the combination of experts with varying degrees of regularization and their adaptive integration. This architecture allows for the simultaneous consideration of robust global dependencies and local data variations, which is particularly important for spatially heterogeneous hydrological processes. Moreover, the stable forecast accuracy observed at different forecast horizons indicates the presence of informative signals preceding the event and associated with moisture accumulation, temperature conditions, and snow cover conditions. This confirms that flood hazard arises from the combined effects of static and dynamic factors, rather than being determined by a single set of features. Importantly, SHAP analysis allows us to determine the relative contributions of individual predictors beyond what can be inferred from traditional correlation analysis, highlighting the value of explainable machine learning for interpreting flood forecasting models. Despite the heterogeneous temporal distribution of observations, the dataset retains balanced class proportions and broad spatial coverage, supporting the consistency of the reported results across the evaluated forecasting scenarios.

Among all predictors, elevation emerged as one of the most influential variables, exceeding the contribution of individual precipitation-related characteristics. This finding is physically consistent, as elevation primarily controls drainage patterns, river network organization, terrain slopes, and runoff accumulation, thereby determining the inherent

susceptibility of an area to flooding. In contrast, the impact of precipitation depends on antecedent hydrometeorological conditions, including snow accumulation, soil moisture, and temperature, and is therefore distributed among multiple complementary predictors aggregated over different time windows prior to the event. Consequently, precipitation-related predictive information is represented collectively, rather than by a single dominant variable, while elevation provides a stable physiographic descriptor across the entire dataset.

These results also have implications for model transferability. Terrain-related predictors, such as elevation, are expected to remain informative across different geographic regions, as they describe relatively invariant physical characteristics. In contrast, the importance of precipitation-related variables is likely to vary across hydroclimatic regimes, particularly in regions dominated by convective or monsoonal precipitation. Therefore, applying the proposed RS-ALOP-ML framework to other climate regions will require regional recalibration of hydrometeorological predictors while maintaining the general physiographic relationships identified from topographic variables.

Beyond its predictive performance, the proposed RS-ALOP-ML framework has broader implications for disaster risk governance and resilience planning. Probabilistic flood hazard forecasts can support evidence-based decision-making by assisting public authorities in prioritizing vulnerable regions, allocating emergency resources, and planning preventive measures before flood events occur. However, the operational deployment of such models also requires careful consideration of governance issues, including transparent data-sharing practices, responsible use of geospatial information, and the potential consequences of model-guided decisions under conditions of uncertainty. In particular, probabilistic forecasts should be regarded as decision-support tools rather than fully autonomous decision-making systems and should complement expert hydrological assessment and institutional emergency management procedures. These considerations are consistent with recent studies emphasizing that resilient governance depends not only on predictive analytics but also on transparent institutional mechanisms, responsible resource allocation, and evidence-based policy frameworks [41].

Limitations of the Study: This study has several limitations. First, the proposed framework integrates satellite, reanalysis, and geomorphometric data with different spatial and temporal resolutions, resulting in an inherently heterogeneous feature space despite the harmonization procedures applied during preprocessing. Second, the experimental evaluation was conducted using flood data from Kazakhstan; therefore, the generalizability of the proposed framework to significantly different hydroclimatic regions has not yet been established. Third, the use of global flood data may underestimate localized and short-lived floods, potentially limiting the model's sensitivity to specific local hydrological conditions.

Another important limitation concerns the relatively small number of independent flood events available for model development. Although the final integrated dataset contains 3387 observations, these observations correspond to only 18 independent flood events; therefore, the effective sample size at the event level is significantly smaller than the sample size at the row level. This is particularly relevant for multiparameter deep learning architectures such as FT-Transformer and TabNet, which typically require a sufficient number of independent training examples to reliably estimate their model parameters. Under current data storage conditions, these models may be more susceptible to overfitting and may not fully realize their potential advantages over traditional machine learning approaches. Therefore, their performance in this study should be interpreted as a comparison under a limited number of events and a paucity of forecasting data, rather than as evidence of their overall inferiority in flood forecasting. To reduce overly optimistic performance estimates, observations related to the same flood event were stored in a single data partition, and an event-based temporal estimation strategy was applied. However, these procedures cannot

fully compensate for the limited number of independent events. Therefore, the presented results should be interpreted within the context of the geographic, temporal, and data availability conditions represented by the current dataset.

Although the RS-ALOP-ML model was evaluated using flood data from Kazakhstan, its reliance on publicly available geospatial, hydrometeorological, and remote sensing data provides a foundation for assessing its applicability to other regions. Therefore, future work will focus on cross-regional and cross-country validation using larger datasets containing a significantly larger number of independent flood events and representing a variety of hydroclimatic conditions, including snowmelt-driven, precipitation-driven, and mixed flood regimes. Such experiments will allow for a more robust assessment of the generalizability of the proposed model and provide more compelling evidence of the comparative effectiveness of deep learning and traditional machine learning methods.

Overall, the obtained results demonstrate that reliable early flood hazard forecasting requires three complementary components: a leakage-safe experimental design, the integration of multimodal geospatial and hydrometeorological information, and robust yet interpretable machine learning models. The proposed RS-ALOP-ML framework demonstrates that carefully designed probabilistic ensemble learning can provide competitive predictive performance while maintaining physical interpretability and operational applicability.

6. Conclusions

This study presents a leakage-proof hybrid machine learning framework for early flood risk forecasting using multimodal spatiotemporal tabular data. Unlike traditional approaches to flood retrospective mapping, the proposed methodology formulates the problem as a true forecasting problem, constraining the learning process to pre-event information and employing event-aware temporal validation to eliminate information leaks. The developed RS-ALOP-ML framework combines multiple logistic regression experts with varying regularization strengths through adaptive probabilistic fusion. This design preserves the interpretability of linear models while improving forecasting stability in heterogeneous environments. Comparative experiments with both classical machine learning algorithms and modern deep learning architectures on tabular data demonstrate that the proposed framework achieves competitive and often superior forecasting accuracy across several evaluation metrics. Beyond forecast accuracy, the main contribution of this work is the development of a reproducible machine learning pipeline for multimodal environmental forecasting. The integration of topographic information, hydrometeorological variables, long-term water dynamics, and satellite-derived spectral indices demonstrates how disparate data sources can be effectively combined within an interpretable computational framework. The proposed evaluation protocol, which eliminates information leakage, further ensures that the reported performance metrics more accurately reflect actual operational deployments rather than optimistic estimates caused by information leakage. The proposed framework remains computationally efficient, requires a relatively small amount of training data compared to deep neural architectures, and produces transparent probabilistic forecasts suitable for decision support systems. These characteristics make the approach attractive for practical early warning applications, where both forecast reliability and model interpretability are important.

Several limitations should be noted. The experimental evaluation was conducted using flood data collected in Kazakhstan, and the number of recorded events remains relatively limited. Therefore, further validation using larger multi-country datasets and more diverse climate conditions is necessary to assess the generalizability of the proposed framework. Future work will explore adaptive multimodal feature learning, uncertainty-aware fore-

casting, graph-based spatial modeling, and transformer-based temporal representations while maintaining the leakage-safety estimation strategy presented in this study.

Overall, this study demonstrates that hybrid interpretable machine learning combined with rigorous leakage-safety testing provides an efficient computational solution for multimodal spatiotemporal forecasting problems and opens a promising direction for future environmental decision support systems.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/computers15090605/s1, Table S1: Data sources, resolution, and potential impact on results; Table S2: Data sources, initial spatiotemporal resolution and matching procedure; Table S3: Statistical diagnostics of RS-ALOP-ML stability by horizons; Table S4: Bootstrap comparison of RS-ALOP-ML with the closest baseline models; Table S5: Spectral indices and their role in flood hazard forecasting.

Author Contributions: Author Contributions: Conceptualization, M.Y. and M.A.; methodology, M.Y. and D.U.; software, A.S. and L.F.; validation, S.T., D.U., Z.M., and A.M.; formal analysis, M.Y. and Z.-G.Y.; investigation, A.S. and L.F.; resources, M.A.; data curation, Z.-G.Y.; writing—original draft preparation, M.Y.; writing—review and editing, S.T., D.U., and A.M.; visualization, A.S. and Z.M.; supervision, M.A.; project administration, M.Y.; funding acquisition, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: In preparing this manuscript, the authors used OpenAI ChatGPT (GPT-5.5) solely for language editing and clarity improvement in the English text. The artificial intelligence tool was not used to generate scientific results, perform data analysis, develop algorithms, interpret results, or formulate conclusions. All scientific content was created, reviewed, and validated by the authors, who bear full responsibility for the final version of the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

AHP	Analytic Hierarchy Process
BSI	Bare Soil Index
ERA5-Land	ECMWF Reanalysis v5 Land
EVI	Enhanced Vegetation Index
F1-score	F1 Measure
FT-Transformer	Feature Tokenizer Transformer
GIS	Geographic Information System
JRC	Joint Research Centre
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
NDMI	Normalized Difference Moisture Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
PR-AUC	Precision–Recall Area Under Curve
ROC-AUC	Receiver Operating Characteristic Area Under Curve
RS	Remote Sensing

RS-ALOP-ML	Remote Sensing Adaptive Linear Opinion Pool Machine Learning
SHAP	SHapley Additive exPlanations
SRTM	Shuttle Radar Topography Mission
TabNet	Attentive Interpretable Tabular Learning Network
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution

References

1. Singh, T.A.; Gautam, S.; Joshi, S.K. Advancements and Challenges in Flood Risk Assessment: A Scientometric Analysis of Global Trends. In *Blue Sky, Blue Water: Strategies for Protecting Air and Water Quality in the 21st Century*; Springer Nature: Cham, Switzerland, 2025; pp. 395–412. [CrossRef]
2. Sett, D.; Bubeck, P.; Hagenlocher, M.; Thieken, A.H. A Systematic Review of Empirical Research on Behavioral Drivers of Households' Flood Risk Adaptation Over Time and Across Regions. *WIREs Water* **2026**, *13*, e70054. [CrossRef]
3. Gudmundsson, L.; Brunner, M.I.; Döll, P.; Fluet-Chouinard, E.; Frolova, N.; Gosling, S.N.; Hirabayashi, Y.; Kireeva, M.B.; Liu, X.; Schmied, H.M.; et al. Past and future change in global river flows. *Nat. Rev. Earth Environ.* **2026**, *7*, 7–23. [CrossRef]
4. Mussina, A.; Makhmudova, L.; Medeu, A.; Ospanova, M.; Abdullayeva, A.; Tursyngali, M.; Zhaksylyk, S.; Rodrigo-Ilarri, J.; Rodrigo-Clavero, M.-E. Integrated Gis-Based Flood Risk Assessment and Satellite-Based Verification in the Zhabay River Basin, Kazakhstan. *Prog. Disaster Sci.* **2026**, *31*, 100694. [CrossRef]
5. Durán, J.D.J.F.; Alcántara-Ayala, I. Recurrent Risk and the Disaster Loop: A Forensic Approach to Urban Flooding. *Int. J. Disaster Risk Reduct.* **2026**, *134*, 106028. [CrossRef]
6. Shahabi, E.; Jafari, M.; Taheri Dehkordi, A. A New Framework for Vertical Accuracy Assessment and Spatial Error Mapping of Global Digital Elevation Models Using ICESat-2. *J. Geovisualization Spat. Anal.* **2026**, *10*, 2. [CrossRef]
7. Pandey, M. Artificial intelligence algorithms in flood prediction: A general overview. In *Geo-Information for Disaster Monitoring and Management*; Springer: Cham, Switzerland, 2024; pp. 243–296. [CrossRef]
8. Rahman, Z.U.; Zhang, M.; Chen, F.; Ullah, S.; Wang, L.; Ahmad, Z.; Baqa, M.F. Spatiotemporal dynamics of flood susceptibility under future precipitation variability, population growth, and land cover change. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *147*, 105193. [CrossRef]
9. Antonopoulos, M.; Juvin-Quarroz, J.; Boucher, O. Hybrid machine learning model for bias correction of UTLS relative humidity against IAGOS observations in ERA5 reanalysis. *Atmos. Chem. Phys.* **2026**, *26*, 4771–4784. [CrossRef]
10. Chai, B. Long-term spatiotemporal dynamics of urban land and surface water bodies in Shanghai over the past 90 years using old maps and dense Landsat time series. *Habitat Int.* **2026**, *168*, 103702. [CrossRef]
11. Ullah, K.; Wang, Y.; Xi, C.; Du, B.; Li, P.; Liu, T.; Hong, S.; Hamed, M.M. Multi-Scenario flood susceptibility projections in the Eastern Hindu kush: Integrating machine Learning, CMIP6, and land use change. *Earth Syst. Environ.* **2025**, *10*, 3793–3812. [CrossRef]
12. Zhang, T.; Zhang, R.; Li, J.; Feng, P. Deep learning of flood forecasting by considering interpretability and physical constraints. *Hydrol. Earth Syst. Sci. Discuss.* **2025**, *29*, 5955–5974. [CrossRef]
13. Kearney, S.P.; Augustine, D.J.; Porensky, L.M.; Peirce, E.S.; Hiestand, M.P.; Derner, J.D. Bringing cross-validation into the real world to evaluate transferability of satellite-based vegetation models. *Sci. Rep.* **2026**, *16*, 9383. [CrossRef] [PubMed]
14. Raieli, S.; Jeanray, N.; Gerart, S.; Vachenc, S.; Altahhan, A. Escaping the forest: A sparse, interpretable, and foundational neural network alternative for tabular data. *npj Artif. Intell.* **2026**, *2*, 14. [CrossRef]
15. Saberian, M.; Samadi, V.; Popescu, I. Probabilistic hierarchical interpolation and interpretable neural network configurations for flood prediction. *Hydrol. Earth Syst. Sci.* **2026**, *30*, 371–399. [CrossRef]
16. Li, Y.; Fang, Z.; Liu, J.; Lu, Z.; Zhou, H.; Yin, W.; Chen, X. A Comparative Study of Urban Pluvial Flood Susceptibility Assessment Based on Multi-Machine Learning Algorithm. *Water Resour. Manag.* **2026**, *40*, 35. [CrossRef]
17. ShravanKumar, S.M.; Karthick, A.; Priya, A.K.; Mohanavel, V.; Muthusamy, S. Remote Sensing and Artificial Intelligence in Flood Prediction: Progress, Challenges, and Prospects. *Arch. Comput. Methods Eng.* **2026**, *33*, 5535–5566. [CrossRef]
18. Song, W.; Guan, M.; Yu, D. SwinFlood: A hybrid CNN-Swin Transformer model for rapid spatiotemporal flood simulation. *J. Hydrol.* **2025**, *660*, 133280. [CrossRef]
19. Amiruddin, H.B.; Arif, S.; Assegaf, M.A.H.; Sakka, S.; Fahruddin, F. Mapping flood susceptibility and settlement exposure using random forest for spatial planning in Barru regency. *Ecol. Eng. Environ. Technol.* **2026**, *27*, 413–429. [CrossRef] [PubMed]
20. Zhang, X.; Guo, D. Interpretable machine learning framework for urban flood susceptibility assessment: A multi-model comparison with spatial heterogeneity analysis in Yancheng. *Sci. Rep.* **2026**, *16*, 21315. [CrossRef] [PubMed]
21. Ademusire, A.J.; Eyinade, J.A. Geospatial machine learning for flood risk assessment in contrasting physiographic environments. *World J. Adv. Eng. Technol. Sci.* **2025**, *16*, 436–446. [CrossRef]
22. Abdolmanafi, A.; Saghafian, B.; Aminyavari, S. Enhancing precipitation intensity estimation using ERA5-land reanalysis with statistical and machine learning approaches. *Results Eng.* **2025**, *26*, 104928. [CrossRef]

23. Zolfaghari, A.A.; Raeesi, M.; Longo-Minnolo, G.; Consoli, S.; Dyck, M. Daily reference evapotranspiration prediction in Iran: A machine learning approach with ERA5-land data. *J. Hydrol. Reg. Stud.* **2025**, *59*, 102343. [CrossRef]
24. Ren, W.; Zhao, T.; Huang, Y.; Honavar, V. Deep learning within tabular data: Foundations, challenges, advances and future directions. *arXiv* **2025**, arXiv:2501.03540. [CrossRef]
25. Jiang, J.P.; Liu, S.Y.; Cai, H.R.; Zhou, Q.L.; Ye, H.J. Representation learning for tabular data: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6488–6508. [CrossRef] [PubMed]
26. Rabbani, S.B.; Medri, I.V.; Samad, M.D. Attention versus contrastive learning of tabular data: A data-centric benchmarking. *Int. J. Data Sci. Anal.* **2025**, *20*, 3069–3091. [CrossRef]
27. Hajji, S.; Krimissa, S.; Abdelrahman, K.; Boudhar, A.; Elaloui, A.; Ismaili, M.; El Bouzekraoui, M.; Essbiti, M.C.; Kahal, A.Y.; Mondal, B.K.; et al. Enhancing flood prediction through remote sensing, machine learning, and Google Earth Engine. *Front. Water* **2025**, *7*, 1514047. [CrossRef]
28. Xu, Y.; Li, H.; Hu, Y.; Zhang, C.; Xu, B. Toward improved deep learning-based regionalized streamflow modeling: Exploiting the power of basin similarity. *Environ. Model. Softw.* **2025**, *188*, 106374. [CrossRef]
29. Gegenleithner, S.; Pirker, M.; Dorfmann, C.; Kern, R.; Schneider, J. Long short-term memory networks for enhancing real-time flood forecasts: A case study for an underperforming hydrologic model. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 1939–1962. [CrossRef]
30. Saberian, M.; Zafarmomen, N.; Panthi, K.; Samadi, V. Unravelling the power of neural networks for flood prediction across complex hydrological systems. *GeoHorizons* **2026**, *1*, gh2025-4. [CrossRef]
31. Singha, C.; Chakraborty, N.; Sahoo, S.; Pham, Q.B.; Xuan, Y. A novel framework for flood susceptibility assessment using hybrid analytic hierarchy process-based machine learning methods. *Nat. Hazards* **2025**, *121*, 13765–13810. [CrossRef]
32. Wang, J.; Jia, Z.; Zhang, W.; Li, X. Method for selecting typical floods based on an unfavorable indicator and flood classification. *Environ. Fluid Mech.* **2025**, *25*, 49. [CrossRef]
33. Desai, A.; Abdelhamid, M.; Padalkar, N.R. What is reproducibility in artificial intelligence and machine learning research? *AI Mag.* **2025**, *46*, e70004. [CrossRef]
34. Semmelrock, H.; Ross-Hellauer, T.; Kopeinik, S.; Theiler, D.; Haberl, A.; Thalmann, S.; Kowald, D. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Mag.* **2025**, *46*, e70002. [CrossRef]
35. Choubin, B.; Jaafari, A.; Henareh, J.; Karimi, O.; Hosseini, F.S. Explainable artificial intelligence (XAI) for interpreting predictive models and key variables in flood susceptibility. *Results Eng.* **2025**, *27*, 105976. [CrossRef]
36. Xu, Q.; Shi, Y.; Bamber, J.L.; Ouyang, C.; Zhu, X.X. Large-scale flood modeling and forecasting with FloodCast. *Water Res.* **2024**, *264*, 122162. [CrossRef] [PubMed]
37. Arahman, N.; Rosnelly, C.M.; Mulyati, S.; Dzulhijjah, W.A.; Halimah, N.; Haikal, R.D.U.; Siddiq, S.; Maulidayanti, S.; Aziz, M.; Ulbricht, M.; et al. Investigation of microplastics in community well water in Banda Aceh, Indonesia: A separation technique using polyethersulfone-poloxamer membrane. *AIMS Environ. Sci.* **2025**, *12*, 53–71. [CrossRef]
38. Waqas, M.; Naseem, A. Artificial intelligence in sustainable industrial transformation: A comparative study of industry 4.0 and industry 5.0. *FinTech Sustain. Innov.* **2025**, *1*, A2. [CrossRef]
39. Tellman, B.; Sullivan, J.A.; Kuhn, C.; Kettner, A.J.; Doyle, C.S.; Brakenridge, G.R.; Erickson, T.A.; Slayback, D.A. Satellite imaging reveals increased proportion of population exposed to floods. *Nature* **2021**, *596*, 80–86. [CrossRef] [PubMed]
40. Dar, A.A.; Hirani, T.; Arora, V.; Kakkar, S. Predictive Analytics for Personalized Debt Management: Leveraging Machine Learning to Provide Actionable Financial Advice. *FinTech Sustain. Innov.* **2025**, *1*, A4. [CrossRef]
41. Charles, D. The Hurricane of Debt Hammering International Trade of Selective CARICOM Member States: Why the Region Needs Structural Reform of Official Development Assistance. *FinTech Sustain. Innov.* **2025**, *1*, A5. [CrossRef]