Journal of Intelligent Monitoring and Control Systems (JIMCS)
DOI: 10.65098/jimcs.01.2026.42.55



REVIEW ARTICLE

Machine Learning for Climate Hazard Modeling: A Critical Review of Methods, Validation, and Research Gaps

Farhad Sadeghi[1], Morteza Pakdaman[2,*]

[1] *Department of Industrial Engineering, Khayyam University, Mashhad 9189747178, Iran*
[2] *Research Institute of Meteorological and Atmospheric Science (RIMAS), Climate Research Institute (CRI), Mashhad, Iran*
** Corresponding Author Email: pakdaman.m@gmail.com*

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received: 30 Aug 2026
Accepted: 09 Oct 2026
Available online: 30 Oct 2026

Online Article Code

ABSTRACT

Machine learning (ML) is increasingly used to map, detect, and forecast climate-related hazards, yet the diversity of targets, datasets, scales, and validation designs limits direct comparison among published results. This critical review synthesizes 49 primary ML application studies addressing floods and stormwater, droughts, temperature extremes, wildfires, and storm-related hazards, while using two review articles and five context sources separately. Tree ensembles and boosting methods are prevalent in tabular susceptibility and risk mapping; support vector methods remain useful for moderate-sized nonlinear problems; and convolutional neural networks, long short-term memory networks, graph, transformer, and hybrid architectures are increasingly applied to imagery and spatiotemporal forecasting. Evidence of superiority is conditional: controlled within-study comparisons sometimes favor an ML model, but headline accuracy or area under the receiver operating characteristic curve values cannot be generalized across regions or validation protocols. The synthesis identifies recurring weaknesses in random data splitting, incomplete treatment of spatial and temporal dependence, class imbalance, possible preprocessing leakage, limited external validation, and weak uncertainty propagation. Physically informed and hybrid approaches may improve credibility and extrapolation, whereas explainability methods reveal model associations rather than causality or compliance with physical laws. Progress toward operational use depends on blocked and out-of-domain validation, calibrated uncertainty, transferable and reproducible workflows, multimodal data integration, and monitoring under nonstationary climate conditions. ML is therefore best positioned as a complementary component of statistical and process-based hazard modeling rather than a universal replacement.

KEYWORDS

Machine Learning; Climate Hazards; Floods; Droughts; Temperature Extremes; Wildfires; Storms; Validation; Uncertainty; Transferability

1. INTRODUCTION

Climate-related hazards arise from interactions among atmospheric, hydrological, land-surface, oceanic, ecological, and human systems. Floods, droughts, heatwaves, cold waves, wildfires, tropical cyclones, severe convective storms, storm surge, and dust transport differ in their governing processes and predictability, but they share practical demands for timely detection, reliable forecasting, and defensible risk assessment. Observed impacts are also shaped by exposure, vulnerability, land-use change, and infrastructure, so a high-performing hazard model is not automatically a complete risk model.

Process-based and statistical models remain foundational because they encode physical knowledge, support scenario analysis, and provide established uncertainty frameworks. They can nevertheless be computationally demanding or difficult to calibrate when observations are heterogeneous, and relationships are strongly nonlinear. Growing archives of remote-sensing products, reanalysis, weather forecasts, station and sensor measurements, geospatial covariates, and socioeconomic data have therefore encouraged complementary machine learning (ML) approaches. Systematic errors persist in weather and climate models (Frassoni et al., 2023), and ML has also been used to post-process ensemble precipitation forecasts (Pakdaman et al., 2022). The value of ML, however, depends on whether the training data represent the deployment setting and whether evaluation prevents information from leaking across space, time, or preprocessing steps.

Previous summaries often emphasize the best metric reported by each study. Such values are not directly comparable when studies use different event definitions, spatial resolutions, forecast horizons, class ratios, sampling schemes, and test sets. This review instead asks four questions: which model families are suited to particular hazard tasks and data structures; how methodological choices affect reported skill; what limits geographic and temporal generalization; and what evidence is needed for credible operational use. The scope uses five consistent categories: floods and stormwater; droughts; temperature extremes, including terrestrial and marine heatwaves and the comparatively sparse cold-wave literature; wildfires; and storm-related hazards, comprising tropical cyclones, severe storms, storm surge, and dust transport. Urban drainage is treated with flood and stormwater applications rather than atmospheric storms. The five-category hazard taxonomy and the principal ML application types used throughout the review are summarized in Figure 1.

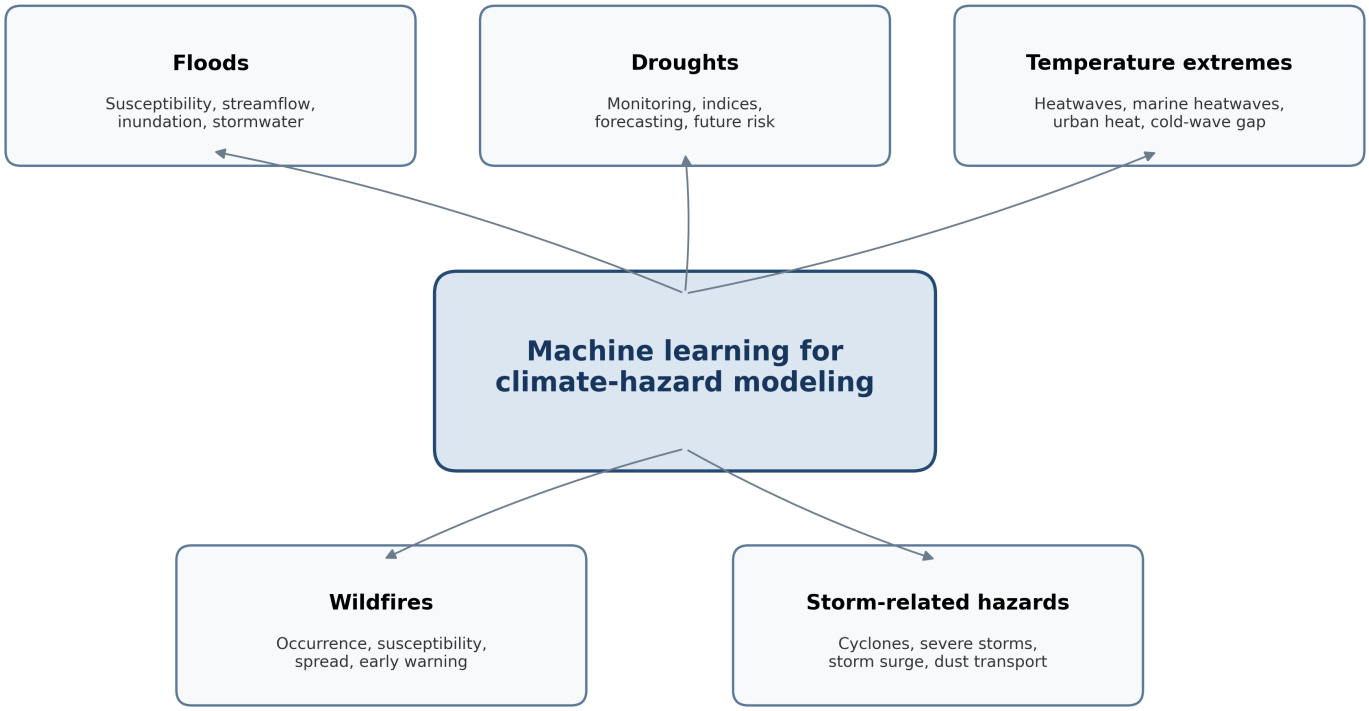


Figure 1 Hazard Taxonomy and Principal ML Application Types Used Consistently Throughout the Review

2. REVIEW METHODOLOGY

2.1 Review Design and Evidence Roles

This article was developed as a structured critical review rather than a systematic review or meta-analysis. The manuscript was based on a pre-existing literature collection assembled during its initial preparation. Because the original literature-identification process was not prospectively documented, the exact database-specific search strings, search dates, numbers of retrieved and duplicate records, and record-level screening decisions cannot be reconstructed reliably. Accordingly, the present review does not claim exhaustive coverage of all publications within a predefined time interval and does not present a Preferred Reporting Items for Systematic Reviews and Meta-Analyses-style study-selection flow.

To strengthen transparency during revision, the retained literature was re-evaluated against explicit scope and eligibility criteria and classified according to its role in the review. The final evidence base comprises 49 primary ML application studies used in the hazard-specific and cross-hazard synthesis, two review articles used for broader methodological or field-level context, and five contextual/background sources used to support climate, forecasting, or hazard interpretation without being counted as primary ML evidence.

2.2 Scope and Eligibility

For the purposes of the revised evidence synthesis, a study was classified as primary evidence when it met the following criteria: (1) it was a scholarly journal article or conference paper; (2) it addressed at least one of the five hazard categories used consistently in this review-floods and stormwater, droughts, temperature extremes, wildfires, or storm-related hazards; (3) ML constituted a substantive component of the hazard-modeling framework rather than an incidental analytical tool; (4) the application involved hazard detection, monitoring, susceptibility or risk mapping, forecasting, reconstruction, spread or pathway prediction, or early warning; and (5) the study reported an empirical application together with quantitative or comparative evaluation.

Sources were not treated as primary ML evidence when they were review articles, broad climatological or hazard assessments without a central ML application, methodological or contextual background papers, or studies outside the defined hazard taxonomy. Studies in which ML was used only as a supporting analytical tool rather than as a substantive component of hazard modeling were retained as contextual evidence when they contributed relevant interpretation. Duplicate reports and non-scholarly material were not included in the analytical evidence base. These criteria were applied during revision to distinguish the studies used in the primary synthesis from review and contextual sources.

2.3 Search Documentation and Scope Definition

The exact database-specific search syntax used during the initial preparation of the manuscript was not retained and therefore cannot be reported retrospectively with sufficient reliability. In particular, the original database coverage, search dates, retrieved-record counts, and duplicate-removal history are unavailable. No reconstructed Boolean query is presented here as though it represented the original search procedure.

For the revised review, the scope of the retained literature was defined using three concept groups consistent with the eligibility criteria: machine-learning methods, climate-hazard categories, and hazard-modeling applications. Method concepts included ML, deep learning, Random Forest (RF), boosting methods, support vector machines (SVMs), neural networks, convolutional and recurrent neural networks (CNNs), transformers, and graph neural networks (GNNs). Hazard concepts covered floods and stormwater, droughts, temperature extremes, wildfires, and storm-related hazards, including tropical cyclones, severe storms, storm surge, and dust transport. Application concepts included detection, monitoring, susceptibility and risk mapping, forecasting, reconstruction, spread or pathway prediction, and early warning. These concepts were used to define and audit the thematic scope of the evidence included in the revised synthesis rather than to reconstruct an undocumented historical database query.

2.4 Data Extraction and Synthesis

For each primary study, the revised review recorded the hazard category, modeling task, study region or application setting, principal data sources, machine-learning method or model family, reported evaluation metrics,

main study-specific findings, and limitations relevant to interpretation. These elements formed the basis of the hazard-specific evidence tables (Tables 2–6). Numerical performance values were retained only when they could be linked to the evaluation design reported in the individual study and were interpreted as study-specific results rather than as directly comparable estimates across the literature.

Synthesis was conducted first within each hazard category and then across hazards. Within-hazard synthesis focused on recurring modeling tasks, data structures, model families, forecast horizons or spatial scales, validation approaches, interpretability, and reported limitations. Cross-hazard synthesis then examined broader patterns in model suitability, validation quality, transferability, data leakage, class imbalance, uncertainty, computational demands, and interpretability. Because the reviewed studies differed substantially in prediction targets, datasets, spatial and temporal scales, class distributions, metrics, and validation procedures, no pooled meta-analysis or cross-study ranking of model performance was performed.

Table 1 Evidence Corpus and Role in the Review

Evidence role	Number	Use in synthesis
Primary ML application studies	49	Hazard-specific and cross-hazard analytical synthesis
Review articles	2	Methodological and field-level background; not counted as primary studies
Context/background sources	5	Climate-model, forecasting, and hazard context; not used as primary ML evidence

Table 2 Primary Evidence for Floods and Urban Stormwater

Study	Application and data	Method	Reported evaluation	Critical interpretation
Amiri et al (2024).	Current/future flood susceptibility; remote sensing, terrain, and scenario inputs	Analytic Network Process (ANP), decision tree, outlier-robust extreme learning machine (ORELM), weighted overlay	Overall flood susceptibility map (FSM) accuracy = 88.3%; future scenario projections	Shows an integrated mapping/scenario workflow; conclusions remain region- and scenario-dependent.
Wahba et al (2024).	Flash-flood susceptibility in Ibaraki, Japan; 11 conditioning factors	Random forest regression and GIS	ROC-AUC reported as 0.9956	Very high internal discrimination should not be interpreted as proven transferability.
Eslaminezhad et al (2021).	Flood-hazard zonation; 15 variables and 300 flood locations	alternating decision tree (ADTree), RF, naïve Bayes tree (NBTree)	AUC comparison	RF led the within-study comparison; inventory size and spatial dependence constrain generalization.
Lee et al (2017).	Urban flood susceptibility in Seoul; environmental predictors	RF and boosted trees	AUC; within-study comparison	RF performed better in the authors' design; the result does not establish universal superiority.
Cheraghi Ghalehsari et al (2020).	Basin-scale susceptibility; environmental and anthropogenic factors	SVM and GIS	Predictive accuracy reported	Demonstrates kernel classification for nonlinear predictors; transfer testing was not emphasized.
Shada et al (2022).	Hourly river-flood forecasting; rainfall and water level	SVM and WSVM; comparison with previously developed WANN	Peak timing and magnitude; error metrics	Wavelet preprocessing improved the same-study comparison, but preprocessing must remain inside validation folds.
Langhammer (2023)	Daily/hourly flood simulation with sensor-network data	SVR	NSE; sensitivity analysis	Daily performance exceeded hourly performance, illustrating scale-dependent skill.
Liu et al (2022).	Regional flood susceptibility across the Belt and Road region	SVM, logistic regression, ensemble	AUC = 0.9342 for ensemble	Large-area mapping is valuable, but heterogeneous regions require spatially explicit validation.
Wang et al (2023).	Urban flood susceptibility with limited inventories	MIC-weighted clustering with AHP-based comprehensive evaluation	Cluster comparison and historical consistency	Unsupervised partitioning reduces label dependence, but cluster validity is not equivalent to predictive probability.
Li et al (2025).	Data-scarce flash-flood classification in Zhejiang	Fractional-order k-means	Validation against 1,643 documented flash-flood sites	Avoids labeled training data; sensitivity to selected variables and cluster number remains.
Li et al (2023).	Urban flood risk and vulnerability in Fuzhou	Risk-vulnerability framework with combined weighting, DBSCAN and K-means	Comparison with TOPSIS; historical agreement	Combines hazard and vulnerability, but method comparisons depend on the adopted risk definition.
Ishak et al (2024).	Flood-prone-area clustering in Bengkulu	k-means and k-medoids	Internal clustering metrics; K-means with k=5 selected	Useful exploratory GIS workflow; independent event-based validation is needed.
Enung et al (2023).	Streamflow forecasts at multiple lead times	MLP with tuned hyperparameters	Skill above 0.90 at <4 h; decline after 6 h	Lead-time degradation is operationally more informative than a single headline score.
Wang et al (2021).	Flood susceptibility in Puyang; 12 conditioning factors	MLP/CART hybrids with FR and IoE	Within-study model comparison	Hybrid MLP combinations led the reported comparison; applicability outside the study area was not established.
Razavi-Termeh et al (2025).	Flood-prone-area mapping in Jahrom	CatBoost with Germinal Center Optimization (GCO), Augmented Artificial Ecosystem Optimization (AAEO) and Water Cycle Algorithm (WCA)	Comparative discrimination; partial dependence plot (PDP)	Optimization improved the local comparison; PDP indicates association, not causality.
Goumghar et al (2025).	Flood-prone mapping in semi-arid Morocco	HistGradientBoosting, XGBoost, CatBoost	Accuracy and error metrics	Boosting was effective for tabular predictors; blocked validation is important in spatially autocorrelated data.

Table 2 (Continued)

Study	Application and data	Method	Reported evaluation	Critical interpretation
Kratzert et al (2019).	Large-sample rainfall-runoff simulation across catchments	LSTM	Cross-catchment comparisons	Supports regional learning, while performance variation among basins underscores transfer limits.
Nemni et al (2020).	Rapid inundation segmentation from SAR imagery	Fully convolutional network	Image-segmentation metrics	Enables rapid mapping; sensor/domain shifts and boundary uncertainty require explicit evaluation.
Wang et al (2020).	Flood susceptibility from geospatial/remote-sensing inputs	CNN	Comparison with SVM benchmark; AUC improvement of 0.022–0.054	CNN improved the reported same-dataset comparison; spatial independence of evaluation remains central.
Ramachandran et al (2024).	Storm-drain placement and reinforcement	RF with QGIS and socioeconomic/physical data	RF accuracy = 96.03%;F1 score = 0.899; planning maps	Belongs to urban flood infrastructure rather than atmospheric storm prediction; evidence is conference-based.

Statements that one model outperformed another were retained only when the comparison was reported within the same study under a shared dataset, prediction target, metric, and evaluation framework. Interpretive statements in the synthesis tables were used to distinguish reported study results from broader methodological implications; limitations such as data leakage or lack of transferability were not attributed to an individual study unless they were supported by the published study design or explicitly acknowledged by the authors. Table 1 summarizes the composition of the final evidence corpus and the role of each source category in the synthesis.

The distribution of the 49 primary ML application studies across the five hazard categories is shown in Figure 2.

3. MACHINE LEARNING IN FLOOD AND STORMWATER MODELING

3.1 Susceptibility Mapping and Tabular Ensembles

Flood-susceptibility studies commonly integrate topographic, hydrological, land-cover, precipitation, geological, and exposure-related variables in tabular form. Tree-based ensembles and boosting methods are well suited to these heterogeneous predictors because they can represent nonlinear interactions while retaining comparatively straightforward measures of predictor importance. Within individual studies, RF achieved stronger discrimination than alternative tree-based models in Maneh and Samalqan County and Seoul (Eslaminezhad et al., 2021; Lee et al., 2017). Other studies extended this approach through ensemble and hybrid modeling. Liu et al. (2022) combined SVM and logistic regression models within an ensemble framework for regional flood-susceptibility assessment, whereas Wang et al. (2021) integrated Frequency Ratio and Index of Entropy measures with multilayer perceptron (MLP) and classification and regression tree (CART) models to construct hybrid susceptibility frameworks. More recent work has explored boosting-based approaches in different regional settings: Razavi-Termeh et al. (2025) enhanced Categorical Boosting (CatBoost) using metaheuristic optimization for flood-prone-area mapping in Jahrom, while Goumghar et al. (2025) compared several boosting algorithms, including Extreme Gradient Boosting (XGBoost), CatBoost, Light Gradient Boosting Machine (LightGBM), and Hist Gradient Boosting, in the Upper Drâa Basin of southern Morocco.

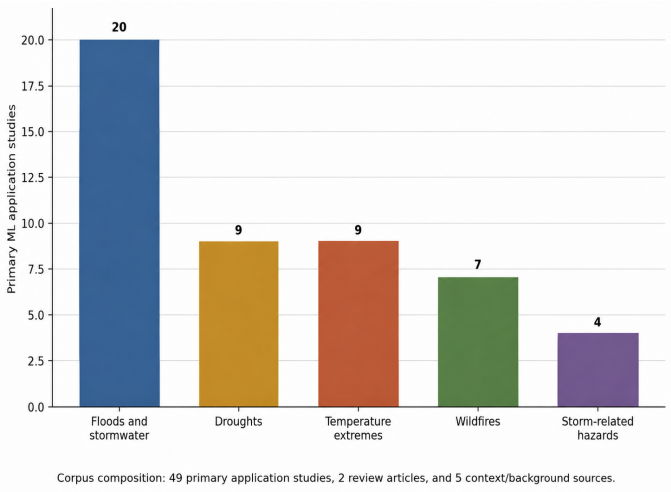


Figure 2 Distribution of the 49 Primary ML Application Studies Across the Consistent Hazard Taxonomy. Urban Stormwater Infrastructure Is Counted with Floods.

Amiri et al. (2024) further extended susceptibility mapping to future conditions by incorporating remote-sensing, geospatial, and scenario-related information, whereas Wahba et al. (2024) reported an area under the receiver operating characteristic curve (ROC-AUC) of 0.9956 for flash-flood susceptibility mapping in Ibaraki, Japan. Collectively, these studies demonstrate the flexibility of ensemble and hybrid methods for tabular flood-susceptibility applications; however, their reported performance remains dependent on the study region, predictor set, sampling strategy, and validation design, and should therefore not be interpreted as evidence of equivalent performance across other basins or future climatic conditions. Kernel-based approaches have also been applied to flood-susceptibility mapping. Cheraghi Ghalehsari et al. (2020) used SVM with geographic information system (GIS)-based geomorphological, hydrological, and physiographic conditioning factors in the Nekaroud Basin, illustrating a non-tree alternative for representing nonlinear flood-susceptibility relationships.

CNN-based approaches provide an alternative when spatial information is represented directly in gridded or image data rather than summarized as tabular conditioning variables. Wang et al. (2020) applied CNN frameworks to flood-susceptibility mapping using geospatial and remote-sensing inputs and reported improved performance relative to statistical comparators within the same study design. This result supports the ability of convolutional architectures to learn spatial patterns from structured environmental data, although the reported advantage remains specific to the dataset and validation framework used in that study. A different application was demonstrated by Nemni et al. (2020), who used a fully convolutional network to segment inundated areas from synthetic aperture radar (SAR) imagery for rapid flood mapping. Although both studies employed convolutional architectures, susceptibility mapping and inundation segmentation represent distinct modeling tasks. Susceptibility models estimate the relative likelihood of flooding from conditioning information and depend strongly on the construction and spatial independence of flood and non-flood samples, whereas image-segmentation performance is influenced by sensor characteristics, image labels, spatial resolution, and boundary delineation. Consequently, performance metrics reported for these two applications should be interpreted within their respective evaluation settings rather than directly pooled or ranked.

3.2 Forecasting, Sequence Models, and Lead Time

Flood-forecasting studies illustrate how temporal resolution and forecast horizon can substantially affect reported model performance. Shada et al. (2022) developed SVM and hybrid Wavelet-SVM (WSVM) models

for hourly flood forecasting in the Achankovil River, using wavelet-based denoising of rainfall and water-level inputs. The WSVM reproduced peak-discharge magnitude and timing accurately and was also reported to perform better than a previously developed wavelet-artificial neural network (WANN) model for the same river. Langhammer (2023) coupled support vector regression (SVR) with hydrometeorological sensor-network data and evaluated flood simulations at both daily and hourly resolutions. Although the daily simulations produced higher mean aggregate performance metrics (Nash-Sutcliffe efficiency (NSE) = 0.880 versus 0.820 for the hourly simulations), the study also showed that daily aggregation simplified hydrograph shape and flood timing and reduced the advantages of high-frequency sensor observations. The hourly resolution therefore provided a more informative balance between predictive skill and temporal detail for operational applications. Enung et al. (2023) similarly demonstrated a strong dependence on forecast horizon in MLP-based streamflow prediction for the Upper Citarum River. Forecasts at 2–4 h leads showed the strongest skill, with NSE and R^2 values around or above 0.90 depending on the input scenario, whereas performance declined progressively at longer horizons, particularly beyond 6 h. The use of areal rainfall inputs also produced slightly better results than station-based rainfall in the reported experiments. Taken together, these studies show that forecast lead time, temporal resolution, peak magnitude, and event timing provide information that cannot be represented adequately by a single aggregate performance score. They also highlight a general validation requirement: data-dependent preprocessing and hyperparameter tuning should be confined to the training framework to prevent information from the evaluation period from entering model development, as discussed further in Section 8.2.

At a larger spatial scale, Kratzert et al. (2019) demonstrated that long short-term memory (LSTM)-based regional rainfall-runoff models can learn both shared and catchment-specific hydrological behavior from large-sample datasets. Using meteorological time series and static catchment attributes across 531 Catchment Attributes and Meteorology for Large-sample Studies basins, the study showed that entity-aware LSTM achieved higher overall NSE than several regionally calibrated and basin-specific hydrological benchmark models under a common evaluation framework, although predictive skill still varied among individual basins. This combination of strong regional performance and basin-to-basin variability illustrates both the potential and the transferability challenge of sequence models. At the literature level, Mosavi et al. (2018) identified hybridization, data decomposition, model ensembles, and optimization as recurring strategies associated with improved ML-based flood prediction. Because those conclusions were synthesized across heterogeneous studies, however, they should not be interpreted as a controlled ranking of model families. Overall, the evidence suggests that model selection for flood forecasting should account not only for predictive accuracy but also for record length, temporal resolution, forecast horizon, regional transferability, and the intended operational use.

3.3 Unsupervised Clustering and Data-Scarce Flood Mapping

Unsupervised clustering provides an alternative to supervised flood-susceptibility models when reliable event inventories are limited or when the objective is to identify spatial groupings in multidimensional geospatial data. Wang et al. (2023) combined clustering with a comprehensive evaluation framework for urban flood susceptibility in Zhengzhou. Maximal Information Coefficient (MIC) was used to weight the relationships between conditioning factors and flooding, while an analytic hierarchy process (AHP)-based evaluation structure was used to compare alternative clustering results; the final framework recommended four susceptibility levels obtained using K-means. This approach reduces dependence on predefined susceptibility-class thresholds and requires relatively limited flood-event information, although the resulting clusters should not be interpreted as calibrated probabilities of flooding.

Li et al. (2025) extended unsupervised flood mapping to explicitly data-scarce settings using a GIS-based fractional-order K-means framework in Zhejiang Province. Five commonly available geospatial variables were clustered without using historical flood occurrences as training labels, and the resulting risk zones were subsequently evaluated against documented flash-flood locations. This separation between label-free model construction and historical-event evaluation is important: it reduces dependence on labeled training inventories without eliminating the need for independent evidence that the resulting clusters correspond to meaningful flood-risk patterns. In a related urban-risk application, Li et al. (2023) integrated Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means within a risk-vulnerability framework for Fuzhou, China. The resulting risk map was consistent with historical flood locations and was compared with both single-clustering and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)-based assessments, illustrating how clustering can be incorporated into a broader risk-evaluation framework rather than treated as a stand-alone predictor.

Ishak et al. (2024) compared K-means and K-medoids for identifying potentially flood-prone areas in Bengkulu Province using rainfall and river-distance information processed within a GIS environment. The published analysis selected K-means with five clusters as the preferred configuration based primarily on internal clustering criteria, including the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, and implemented the resulting clusters in a GIS-based interface. Such internal metrics are useful for evaluating cluster compactness and separation, but they do not demonstrate predictive accuracy against independent flood events. More generally, unsupervised methods can support exploratory zoning and planning where labels are sparse, but their outputs remain sensitive to predictor selection, preprocessing and scaling, the number of clusters, and the criteria used to assign hazard meaning to each cluster.

Storm-drain planning is included within the flood and stormwater category because its objective concerns urban drainage infrastructure rather than atmospheric storm prediction. Ramachandran et al. (2024) combined QGIS with an RF classifier and information on rainfall, soil characteristics, population growth, and flood sensitivity to support storm-drain placement and reinforcement. The study illustrates the use of machine learning for infrastructure-oriented decision support; however, its conference-based planning application should be distinguished from independent validation of flood-occurrence, inundation, or forecasting models. Table 2 summarizes the primary evidence for flood and urban stormwater applications, including the reported evaluation and the principal methodological interpretation of each study.

Across flood applications, model suitability depends strongly on the alignment among model architecture, data structure, prediction target, spatial or temporal scale, and validation design. Tree-based ensembles provide effective baselines for tabular susceptibility mapping, CNN-based approaches are suited to gridded and image-based spatial information, and sequence or kernel-based models support temporal forecasting across different lead times. A recurring limitation across these applications is the limited use of independent spatial, temporal, and cross-basin evaluation, which constrains conclusions about geographic and operational generalization despite continued advances in model complexity.

4. MACHINE LEARNING IN DROUGHT MODELING

Drought applications differ substantially in their target definition, including meteorological indices, soil-moisture deficits, hydrological drought, agricultural drought, and composite risk; consequently, performance reported for these tasks should not be treated as directly comparable. Wang et al. (2022) integrated Moderate Resolution Imaging Spectroradiometer (MODIS), Global Precipitation Measurement (GPM), meteorological, and terrain-related information within an RF framework for grassland drought monitoring in Inner Mongolia. Although the model showed a strong correlation with six-month Standardized Precipitation Evapotranspiration Index (SPEI-6) during training (R = 0.9706), the correlation decreased to 0.6387 on the test set, while the reported consistency rate between model-derived drought classes and SPEI reached 87.9%. The train-test difference illustrates why independent evaluation is more informative than training performance alone. Huang et al. (2025) addressed a different problem by developing a meteorological-drought risk framework for mainland China based on 13 hazard, exposure, and vulnerability indicators. Artificial neural network (ANN) and RF models were compared for historical risk assessment, after which the better-performing RF framework was used to project drought risk from 2021 to 2100 under multiple Shared Socioeconomic Pathway (SSP) scenarios. Such future assessments extend ML beyond historical monitoring, but their outputs inherit uncertainty not only from the fitted model but also from climate

forcing, socioeconomic assumptions, and scenario selection.

Index construction and time-series forecasting introduce different methodological considerations. Alshahrani et al. (2024) developed a support vector machine-based drought index (SVM-DI) for northern Pakistan in which an SVM-based X-bar control-chart framework and station-specific weighting were used to aggregate regional precipitation information. Compared with the Regional Standard Precipitation Index (RSPI) benchmark, the resulting SVM-DI showed lower variability in its correlations with station-level Standardized Precipitation Index (SPI) values and was designed to better accommodate extreme values in regional precipitation data. In forecasting applications, Lee et al. (2013) demonstrated a clear dependence between drought accumulation scale and useful forecast horizon: MLP forecasts were strongest at approximately one month for SPI (3), remained useful for roughly one to three months for SPI (6), and retained skill to about five months for SPI (12). Ali et al. (2017) independently applied an MLP neural network to monthly SPEI series from 17 climatological stations in northern Pakistan and evaluated predictions using mean absolute error (MAE), correlation, and root mean square error (RMSE). Together, these studies show that drought-index construction, accumulation scale, regional aggregation, and forecast horizon materially affect model interpretation and should be reported explicitly rather than reduced to a single performance value.

More recent studies have expanded drought modeling toward explainable and explicitly spatiotemporal architectures. Li et al. (2025) used XGBoost to predict hydrological drought categories in the Huaihe River Basin and integrated SHapley Additive exPlanations (SHAP) to examine spatially and seasonally varying predictor contributions, achieving an overall classification accuracy of 79.9%. These attributions describe how the fitted model uses meteorological, soil-moisture, evapotranspiration, and large-scale climate information and should not be interpreted as evidence of physical causality. Dai et al. (2025) constructed an inferred network among drought centroids and incorporated it into a spatio-temporal graph convolutional network to predict root-zone soil moisture (RZSM)-based agricultural drought indices up to four weeks ahead across the Yangtze River Basin. The model demonstrated useful test performance, although the inferred graph structure and its transferability to other basins remain important evaluation considerations. Duan et al. (2025) addressed long-term soil-moisture data gaps using an LSTM framework across nine major Chinese basins, reporting test performance of $R^2 = 0.928$ and RMSE = 0.020 $m^3\ m^{-3}$ for soil-moisture reconstruction. Expected-Gradients-based explainability was then used to examine the relative contributions of hydrometeorological variables to reconstructed soil-moisture dynamics. This workflow illustrates the value of deep learning for data reconstruction and interpretation, while also emphasizing that uncertainty introduced during gap filling can propagate into subsequent drought analysis.

Table 3 Primary Evidence for Drought Modeling

Study	Application and data	Method	Reported evaluation	Critical interpretation
Wang et al (2022).	Grassland drought monitoring; MODIS, GPM, meteorological-station and terrain data; SPEI6 reference	RF	Training R = 0.9706; test R = 0.6387; consistency 87.9%	The train-test gap is more informative than the training correlation alone.
Huang et al (2025).	Historical and future meteorological drought risk in mainland China; 13 hazard, exposure, and vulnerability indicators	ANN and RF; best-performing RF used for future projections	Historical ANN/RF comparison; future drought-risk projections for 2021–2100 under four SSP scenarios	Integrates hazard, exposure, and vulnerability; future estimates additionally inherit climate-model, socioeconomic, and scenario uncertainty.
Alshahrani et al (2024).	Regional drought index for northern Pakistan	SVM-based regional drought index using an X-bar control-chart framework	Correlation and coefficient-of-variation comparison with RSPI	SVM-DI showed greater regional consistency than the RSPI benchmark; index construction should not be directly compared with drought-forecasting models.
Lee et al (2013).	SPI forecasting in South Korea	MLP	Skill by SPI scale and lead time	Longer accumulation scales were more predictable; short-scale skill decayed rapidly with lead time.
Ali et al (2017).	SPEI forecasting across 17 stations in Pakistan	MLP	MAE, correlation coefficient (R), and RMSE across 17 stations	Multi-station evaluation supports performance within the study region, but geographic transfer beyond northern Pakistan was not evaluated.
Li et al (2025).	Hydrological drought prediction and seasonal predictive-feature analysis	XGBoost with SHAP	Overall classification performance = 79.9%; precision, recall, Heidke Skill Score (HSS), and SHAP attribution	Seasonal attributions reveal changing associations, not causal effects.
Dai et al (2025).	Up-to-four-week RZSM-based agricultural drought prediction in the Yangtze River Basin	Spatiotemporal graph convolutional network	Up-to-4-week forecasts; test RMSE = 0.45 and $R^2 = 0.66$	Graph structure can represent spatial dependence; graph construction and out-of-basin transfer require scrutiny.
Duan et al (2025).	Long-term soil-moisture drought assessment across China	LSTM gap filling with Expected-Gradients explainable artificial intelligence (XAI)	Test $R^2 = 0.928$; RMSE = 0.020 $m^3\ m^{-3}$; MAE = 0.015 $m^3\ m^{-3}$	Supports large-scale temporal modeling, but missing-data reconstruction can propagate uncertainty.
Tanriverdi and Batmaz (2025)	U.S. drought prediction using soil and meteorological data	Boosting, CNN/LSTM/Transformer, hybrid, and ensemble models	Accuracy, precision, recall, F1, and training-time comparison across three regional clusters	Deep-learning models achieved higher within-study accuracy at substantially greater computational cost; performance also varied by regional cluster, limiting a single global model ranking.

Tanriverdi and Batmaz (2025) provided a broader comparison of boosting, recurrent, convolutional, Transformer, hybrid, and ensemble models for U.S. drought prediction. Attention-LSTM and Transformer architectures achieved the strongest performance in different regional clusters, whereas simpler boosting approaches required substantially less computation. This comparison highlights a recurring operational trade-off: improvements in predictive performance may be accompanied by greater computational cost and model complexity. Table 3 summarizes the primary drought-modeling studies and highlights differences in targets, model families, evaluation strategies, and methodological limitations.

Taken together, these studies show that drought-model performance depends strongly on target definition, spatial region, drought accumulation scale, forecast horizon, input-data structure, and validation design. Multi-source inputs and spatiotemporal architectures can provide useful additional information, but the reviewed evidence does not support a single universally superior algorithm. Evaluation should therefore report performance by region, season, drought category, and lead time where relevant, preserve chronological separation between model development and testing, and explicitly consider uncertainty arising from observational products and future climate scenarios

5. MACHINE LEARNING IN TEMPERATURE EXTREMES

The primary ML evidence in this category is concentrated on terrestrial and marine heatwaves, whereas dedicated ML applications for cold-wave prediction remain comparatively sparse. Cold waves are therefore retained within the broader temperature-extremes taxonomy, but the imbalance in the available evidence is treated as a research gap rather than implying equivalent coverage of hot and cold extremes. Peterson et al. (2013) provides broader climatological context for heat- and cold-wave variability, while Shan et al. (2024) contributes a daily-scale framework for identifying compound drought-heatwave events; neither is treated here as

primary ML evidence. Zandler and Samimi (2024), although incorporating RF and Boruta-based variable-importance analysis into an urban-tree cooling assessment, is used primarily as remote-sensing and urban-heat context because ML is a supporting rather than central hazard-modeling component.

For terrestrial heatwaves, ML applications span seasonal prediction and local urban downscaling. Kan et al. (2025) compared five classifiers for heatwave-occurrence prediction across Sweden at lead times of one to five months. Balanced RF provided the most stable performance, with balanced accuracy of approximately 0.77 across the investigated lead times. SHAP analysis further indicated that slowly varying land-surface variables, including runoff and soil water, became increasingly important in model outputs at longer leads, illustrating how the relative value of atmospheric and land-surface predictors can change with forecast horizon. At the urban scale, Gkirmpas et al. (2025) combined a mesoscale meteorological model (MEMO), low-cost sensor observations, urban spatial features, and RF regression to predict street-level temperatures during a heatwave in Thessaloniki. Modeling the residual between observed and MEMO-simulated temperatures provided a practical way to incorporate micro-environmental spatial information that was not fully represented by the mesoscale model. These two studies address fundamentally different prediction scales and therefore should not be compared through a common performance ranking.

Short- and medium-range heatwave forecasting further illustrates the dependence of model performance on lead time, climatic setting, and event rarity. Bhoopathi et al. (2024) compared SVR and XGBoost for seven- and fifteen-day forecasts across four Indian temperature zones. SVR better captured temperature peaks at the seven-day lead in the hottest zones, whereas XGBoost showed a relative advantage at the fifteen-day lead in Zone 1; both models performed less effectively in the cooler Zones 3 and 4. Choudary et al. (2025) compared nine ML and deep-learning approaches for heatwave-event prediction in Tamil Nadu and highlighted the effect of severe class imbalance. In particular, an LSTM model achieved high overall accuracy while failing to identify the minority heatwave class, with precision, recall, and F1-score reported as zero. This provides a clear example of why accuracy alone can be misleading for rare-event prediction and why class-specific metrics are required. At the global scale, Lopez-Gomez et al. (2023) trained convolutional neural weather models to predict extreme temperature anomalies from one to 28 days ahead using loss functions designed to emphasize extremes. Their best extreme-focused model showed positive regressive skill relative to the European Centre for Medium-Range Weather Forecasts (ECMWF) subseasonal-to-seasonal control forecast after two weeks, although ECMWF remained superior in the medium range and the long-lead regressive advantage did not consistently translate into superior binary classification skill. The result therefore supports a conditional within-study advantage rather than a general replacement of physics-based numerical weather prediction (NWP).

Marine temperature extremes extend the evidence to both prediction and future climate assessment. He et al. (2024) combined an LSTM model with Expected Gradients to quantify the contributions of meteorological inputs to marine-heatwave predictions and used additional clustering and decomposition analyses to examine recurring model-response patterns. Such attribution describes how the fitted model responds to its inputs and can support hypotheses about associated processes, but it does not by itself establish physical causality. Jiang et al. (2025) combined RF, XGBoost, LightGBM, and CatBoost with SHAP analysis to examine environmental factors associated with marine-heatwave variability in global coral-reef regions. The ML models showed strong within-study explanatory performance, while temperature variability and ocean-heat-content variability emerged among the most influential predictors. In a separate scenario-based component of the same study, Optimum Interpolation Sea Surface Temperature (OISST) and Coupled Model Intercomparison Project Phase 6 (CMIP6) data were used to assess historical and future marine heatwaves and marine cold spells. Under SSP5–8.5, the projected area ratio of marine heatwaves to marine cold spells reached approximately 196 by 2070. The ML-based driver analysis and the CMIP6-based future projection should therefore be interpreted as complementary components of the study rather than as evidence that the ML models themselves generated the future climate projections.

Table 4 Primary ML Evidence for Temperature Extremes

Study	Application and data	Method	Reported evaluation	Critical interpretation
Kan et al (2025).	Seasonal heatwave-occurrence forecasting in Sweden; remote-sensing, atmospheric, and landscape predictors	Balanced RF with SHAP	Balanced accuracy ≈ 0.77 across 1–5 month leads; heatwave recall decreased from 0.87 to 0.81	Class balancing and long lead time are strengths; calibration and spatial transfer still require testing.
Gkirmpas et al (2025).	Street-scale temperature during a heatwave	MEMO model plus RF residual correction	Local temperature error metrics	Hybrid downscaling uses urban morphology effectively; evidence is event- and city-specific.
Marina et al (2026).	European heatwave detection and attribution	Multivariate Autoencoder Flow-Analogue Method (MvAE-AM) with SHAP	Reconstruction comparison with the classical multivariate analogue method; attribution analysis of major European heatwaves	SHAP relationships may be physically consistent but do not prove causality or governing-law compliance.
He et al (2024).	Sea surface temperature (SST) prediction and marine-heatwave pattern analysis at 13 coastal stations in China	LSTM with Expected Gradients, clustering, and Additive Decomposition	Mean squared error (MSE) and RMSE across repeated data partitions; attribution and pattern analysis	Attribution describes model sensitivity; physical mechanisms require independent process evidence.
Bhoopathi et al (2024).	7- and 15-day heatwave forecasts across Indian temperature zones	SVR and XGBoost	7- and 15-day maximum-temperature/heatwave forecast performance by temperature zone	Model ranking changed with lead time and climate zone, arguing against universal rankings.
Choudary et al (2025).	Heatwave-event prediction in India under class imbalance	Nine machine-learning and deep-learning models including GNNs and Transformers	Accuracy, ROC-AUC, precision, recall, and F1; LSTM accuracy = 0.9824 but heatwave precision/recall/F1 = 0	Rare-event handling is central; accuracy alone would obscure missed extremes and false alarms.
Lopez-Gomez et al (2023).	Global extreme-heat forecasts to 28 days	Extreme-focused neural weather models with transfer learning and custom loss functions	Matched-lead comparison with ECMWF; positive regressive skill relative to the control after 2 weeks	Long-lead regressive gains were conditional; ECMWF remained superior in the medium range, and the advantage did not consistently extend to extreme-event classification.
Buschow et al (2024).	Drivers and interactions in European heatwaves	Logistic principal component analysis (PCA), neural-network regression, and Shapley values/interactions	Nonlinear model explained ≈70% of reduced heatwave variability	Interaction analysis identifies predictive associations between circulation and land-surface variables.

Table 4 (Continued)

Study	Application and data	Method	Reported evaluation	Critical interpretation
Jiang et al (2025).	Marine heatwave-driver analysis and historical/future MHW-MCS assessment across global coral-reef regions; OISSTv2.1, CMIP6 and environmental variables	RF, XGBoost, LightGBM and CatBoost with SHAP	ML model evaluation with train/test separation and cross-validation; SHAP attribution; scenario analysis to 2070	ML was used to analyze MHW-related predictor importance, whereas future marine heatwaves (MHWs) and marine cold spells (MCSs) projections were derived separately from climate data; SHAP attribution should not be interpreted as physical causality.

Interpretability and attribution studies also demonstrate both the value and the limitations of explainable ML. Marina et al. (2026) combined deep autoencoders with an analogue-based framework for European heatwave reconstruction and attribution and used SHAP to interpret the learned relationships. The reported feature relationships were broadly consistent with physical understanding, but such consistency should not be interpreted as proof that the model enforces governing physical laws. Buschow et al. (2024) used logistic principal-component analysis, neural-network regression, and Shapley values and interactions to quantify the information on European heatwaves contained in circulation and land-surface predictors. Their nonlinear framework explained approximately 70% of the reduced heatwave variability, with 500-hPa geopotential and upper-level soil moisture providing the largest individual contributions. Importantly, the authors explicitly framed these quantities as predictive associations rather than causal effects. Welandawe et al. (2025) is retained as a review source for broader context on ML-based marine-heatwave prediction and is not counted as a primary application study. Table 4 summarizes the primary ML evidence for temperature extremes, including terrestrial and marine heatwave applications.

Across temperature-extreme applications, reported model performance is strongly conditioned by forecast horizon, spatial scale, class prevalence, and the specific prediction target. Balanced tree ensembles can provide stable seasonal classification under severe event imbalance, whereas neural and hybrid architectures extend modeling to global forecasts, marine heatwaves, reconstruction, and attribution tasks. However, high aggregate accuracy does not necessarily imply reliable detection of rare heatwave events, and improvements relative to physical forecasting systems can depend on both lead time and evaluation metric. Explainability methods such as SHAP, Expected Gradients, and Shapley interactions provide useful information about model associations and sensitivities but do not establish physical causality. Robust evaluation therefore requires class-specific metrics, temporally and geographically independent testing, and a clear distinction among prediction, reconstruction, attribution, and climate-scenario projection.

6. MACHINE LEARNING IN WILDFIRE MODELING

Wildfire occurrence and susceptibility models commonly combine meteorological, vegetation, terrain, accessibility, and human-activity variables, but their predictive behavior can vary substantially across land-use classes and geographic settings. Caiado and Marques (2025) compared Logistic Regression, RF, and XGBoost for wildfire-occurrence prediction across four Portuguese districts and several land-use categories. With Synthetic Minority Oversampling Technique (SMOTE) used to mitigate class imbalance, performance still differed markedly by land use: brush fires were the most predictable, whereas settlement fires remained more difficult because of greater human-related variability. The study also showed that resampling improved recall at the expense of precision, illustrating why class-specific metrics are more informative than overall accuracy alone. Singha et al. (2024) extended susceptibility modeling to a multimodal climate-change context by integrating 44 topographic, climatic, hydrological, vegetation, radar, forest-health, and anthropogenic factors across ten neural, boosting, RF, and optimized-RF models in the Similipal Tiger Reserve, India. The resulting models showed generally strong within-study discrimination and were subsequently used to examine susceptibility under SSP2–4.5 and SSP5–8.5 future-climate scenarios. Such projections broaden the role of ML from present-day susceptibility mapping to scenario assessment, but they also inherit uncertainty from both the predictive model and the future climate inputs.

Tran et al. (2024) evaluated XGBoost and three metaheuristically optimized variants for wildfire-susceptibility mapping on Hawai'i Island using 14 conditioning factors and multiple discrimination metrics. Optimization produced only a modest improvement over the baseline XGBoost model: AUC increased from 0.9164 to 0.9269 for the best-performing Black Widow Optimization-XGBoost configuration. SHAP analysis identified road proximity, annual temperature, and elevation among the most influential model predictors. The limited magnitude of the AUC improvement suggests that gains from additional optimization should be considered alongside the added tuning complexity and should be confirmed under spatially independent evaluation before being interpreted as evidence of broader generalization. Data scarcity motivates a different class of methods. Ma et al. (2024) evaluated a transductive support vector machine (TSVM) for forest-fire susceptibility assessment in Dayu County, China, under progressively larger labeled-sample scenarios. The TSVM consistently outperformed supervised benchmarks such as RF, SVM, and logistic regression when only a small number of labeled fire locations were available, demonstrating the value of incorporating unlabeled samples in data-scarce settings. However, because the transductive approach exploits the specific unlabeled samples available during model development, its within-study advantage does not by itself establish transferability to a geographically distinct region; independent geographic testing or adaptation would still be required.

Dynamic wildfire spread and early-warning applications require models that represent both spatial structure and temporal evolution. Marjani et al. (2023) developed FirePred, a hybrid multi-temporal CNN trained on 177 wildfire events in British Columbia, Canada, and reported a best F1-score of 94% under the selected parameter configuration. Additional evaluations using wildfire events in Alaska and Nova Scotia showed that performance was sensitive to regional conditions, while uncertainty analysis indicated that prediction uncertainty was greatest near fire boundaries. These findings provide useful evidence of the importance of geographic transfer testing rather than relying only on within-region performance. Jiang et al. (2023) proposed WFNet, a hierarchical CNN that encodes evolving fire state and multimodal environmental conditions through a spread spatiotemporal distribution field. The framework demonstrated competitive predictive accuracy and computational efficiency, but its training and evaluation relied primarily on wildfire-spread simulations generated using the Farsite model in California, supplemented by comparison with a real wildfire case. Consequently, its computational advantages and ability to represent continuous spread dynamics are promising, while broader operational credibility still requires validation across independently observed fires and different fire regimes.

Bhowmik et al. (2023) developed a multimodal early-warning framework based on a U-Convolutional LSTM architecture and a database combining historical wildfire, environmental, meteorological, sensor, and geological information. The model was trained using data from 2012–2017 and evaluated on 2018, where the authors reported prediction accuracy above 97% under their temporal holdout design. A retrospective evaluation across the 2018–2022 fire seasons also identified 85.7% of wildfires larger than 300,000 acres. The use of a temporally later evaluation period provides stronger evidence than a purely random split, although changing sensors, land conditions, climate regimes, and wildfire behavior may still reduce performance when the model is deployed outside the historical development distribution. Table 5 summarizes the primary evidence for wildfire occurrence, susceptibility, spread prediction, and early-warning applications.

Across wildfire applications, model requirements change substantially with the prediction target. Tree-based and boosting methods are effective for occurrence and susceptibility mapping with heterogeneous environmental predictors, whereas semi-supervised approaches can reduce dependence on large labeled fire inventories and convolutional or recurrent architectures can represent dynamic spread and early-warning processes. The

Table 5 Primary Evidence for Wildfire Modeling

Study	Application and data	Method	Reported evaluation	Critical interpretation
Caiado and Marques (2025)	Occurrence prediction across Portuguese districts and land uses	Logistic regression, RF, XGBoost; SMOTE	F1, precision, and recall by district and land-use class; brush-fire F1 > 0.93	Performance varied substantially by land use; SMOTE should be applied only within the training procedure to avoid leakage into evaluation data.
Singha et al (2024).	Current/future susceptibility in India; 44 multimodal factors	ANN, boosting, RF and optimized RF	ROC-AUC > 0.88 and Kappa > 0.62 across models; future susceptibility mapped under SSP2–4.5 and SSP5–8.5; Boruta/SHAP attribution	Scenario maps combine predictive and scenario uncertainty and should not be treated as deterministic forecasts.
Tran et al (2024).	Wildfire susceptibility on Hawai'i Island	XGBoost and metaheuristic hybrids; SHAP	AUC = 0.9164–0.9269; sensitivity, specificity, predictive values, and average precision/PR evaluation	Optimization produced a modest local gain; SHAP provides model attribution rather than causal proof.
Ma et al (2024).	Small-sample forest-fire susceptibility in Dayu County, China	Transductive SVM vs supervised models	Accuracy across labeled-sample sizes	Semi-supervised learning improved performance under limited labels, but the within-study advantage does not establish geographic transferability.
Marjani et al (2023).	Multi-temporal wildfire-spread prediction in British Columbia, with transfer tests in Alaska and Nova Scotia	Hybrid multi-temporal CNN	F1 = 94%; tests in other regions; uncertainty	Regional sensitivity and boundary uncertainty show why headline F1 is insufficient.
Jiang et al (2023).	Continuous spatiotemporal spread prediction	Hierarchical CNN (WFNet)	Training/testing primarily on Farsite-simulated wildfire spread; comparison with Farsite and cellular automata on a real-fire case; predictive accuracy and computational efficiency	Strong simulation-based performance and efficiency do not by themselves establish operational transfer to independently observed wildfire spread.
Bhowmik et al (2023).	Multimodal wildfire prediction and early warning using a California wildfire database (> 37 million data points)	U-Convolutional LSTM	>97% accuracy on 2018 after training on 2012–2017; 85.7% of wildfires >300,000 acres identified in retrospective 2018–2022 evaluation	The later-year evaluation is more informative than a random temporal split, but evidence remains geographically concentrated in California and does not establish robustness under future sensor, climate, or fire-regime shifts.

Table 6 Primary Evidence for Storm-Related Hazards

Study	Application and data	Method	Reported evaluation	Critical interpretation
Youn et al (2026).	Tropical-cyclone pressure-field reconstruction from wind	RF regression	Comparison with parametric and data-driven baselines; ERA5-based evaluation and satellite-wind application	Captures asymmetric pressure structure and land-sea contrasts; performance depends on wind-input quality, and cross-basin transfer was not independently established.
Mecikalski et al (2021).	Severe-warning nowcasting for 2,004 tracked storms using radar, 1-min GOES-14 and lightning predictors	RF	Storm-level 70/10/20 train-validation-test split; CSI, false alarm ratio (FAR), POD, HSS, Peirce Skill Score (PSS) and AUC; multimodal CSI = 0.747 versus 0.636 for radar-only	Storm-level independence strengthens the comparison; multimodal predictors improved skill, but the dataset was restricted mainly to May daytime Great Plains storms.
Xie et al (2023).	Storm-surge forecasting from wind fields and water levels	CNN-based deep-learning surrogate	RMSE and correlation across forecast horizons up to 72 h; evaluation primarily against Finite Volume Community Ocean Model (FVCOM) -generated storm-surge simulations	Fast surrogates can support warning systems but require testing outside simulated/training storm regimes.
Yarmohamadi et al (2025).	Dust-storm pathway prediction from MERRA-2 variables	CNN-LSTM and ConvLSTM with RF-based contextual feature selection	6-, 12-, 18-, and 24-h forecasts; accuracy, precision, recall, F1 and Kappa; CNN-LSTM showed higher Kappa than ConvLSTM	Contextual meteorological variables improved both architectures, but performance derived from MERRA-2-based inputs requires validation under independent observational and operational conditions.

reviewed evidence also shows that strong within-study metrics do not guarantee operational generalization: performance varies across land-use classes and regions, simulated spread data may not reproduce all characteristics of observed fires, and temporal holdouts remain vulnerable to future changes in sensors, climate, vegetation, and fire regimes. Evaluation should therefore combine class-specific metrics with geographically and temporally independent testing, explicit uncertainty assessment, and clear distinction between performance demonstrated on observed events and that obtained from simulated fire-spread data.

7. MACHINE LEARNING IN STORM-RELATED HAZARDS

Storm-related ML applications considered in this review address physically distinct prediction targets. Tropical-cyclone pressure reconstruction estimates storm structure from wind information, severe-storm nowcasting classifies short-term convective risk, storm-surge models predict coastal water-level response, and dust-pathway models represent atmospheric transport. Because these tasks differ substantially in their input data, spatial and temporal scales, predictands, and evaluation metrics, their reported performance should be interpreted within each application rather than combined into a common ranking.

For tropical cyclones, Youn et al. (2026) developed an RF regression framework to reconstruct pressure fields from wind information, addressing the inability of radially symmetric parametric formulations to represent asymmetric storm structure and land-sea contrasts. Trained primarily on ERA5 reanalysis data, the RF model was compared with linear, decision-tree, boosting, and parametric baselines and showed lower errors and stronger spatial representation, particularly for intense cyclones and within the gale-force wind radius. A preliminary application using satellite-derived wind fields further demonstrated the potential for near-real-time pressure reconstruction, but performance varied with the quality and coverage of the satellite wind inputs, emphasizing the importance of observational-domain transfer in operational use. Severe-storm nowcasting presents a different validation problem. Mecikalski et al. (2021) combined radar observations, one-minute Geostationary Operational Environmental Satellite-14 (GOES-14) satellite fields, and lightning information within an RF framework using 2,004 tracked storms. Complete storm tracks were separated among training, validation, and test datasets, avoiding leakage of time steps from the same storm across evaluation subsets. The multimodal predictor set improved severe-warning skill relative to the radar-only model, with the mean critical success index (CSI) increasing from 0.636 to 0.747 and probability of detection (POD) from 0.66 to 0.72. Importantly, the reported 95% value in this study refers to the confidence level of the performance improvement rather than to a 95% forecast accuracy. This design illustrates the value of both multimodal observations and

Table 7 Cross-Method Comparison for Climate-Hazard Applications

ML family	Typical applications and inputs	Strengths	Interpretability and burden	Validation and generalization concerns
Decision trees and RF	Tabular susceptibility, occurrence, reconstruction; mixed geospatial predictors	Robust baseline; nonlinear interactions; moderate data needs	Moderate interpretability; low-to-moderate computation	Spatial autocorrelation; unstable importance under correlated predictors; limited extrapolation
Boosting: XGBoost, CatBoost, histogram methods	Tabular mapping and classification	High predictive flexibility; handles complex interactions	SHAP/PDP available; tuning burden moderate	Over-tuning; class imbalance; local gains may not transfer
SVM/SVR	Moderate-sized classification, indices, and short-term forecasting	Effective in high-dimensional nonlinear settings	Kernel-dependent; moderate burden at scale	Sensitive to scaling and tuning; external calibration often limited
MLP/ANN	Tabular regression and forecasting	Flexible nonlinear approximation	Lower intrinsic interpretability; moderate burden	Needs representative samples; forecast skill degrades with horizon
CNN	Imagery, gridded fields, inundation and spread	Learns spatial features directly	High burden; attribution methods are approximate	Domain/sensor shift; boundary uncertainty; spatial leakage
LSTM and sequence models	Streamflow, drought, wildfire evolution	Represents temporal memory and variable lags	High data and compute needs	Chronological validation; nonstationarity; long-lead degradation
Transformers and neural weather models	Large spatiotemporal forecasting datasets	Long-range dependencies and transfer learning	Very high data/compute needs	Benchmark dependence; calibration; out-of-period extremes
Graph neural networks	Networked drought centers or irregular spatial domains	Represents explicit connectivity	Graph construction adds assumptions	Graph leakage; topology transfer; sensitivity to adjacency definition
Hybrid and process-coupled models	Residual correction, simulation-based surrogates, signal decomposition, optimized and coupled forecasts	Can combine data efficiency, speed, and physical structure	Complex implementation and diagnosis	Validate each component; enforce-not merely explain-physical constraints

storm-level independence when evaluating rare convective hazards.

For coastal storm surge, Xie et al. (2023) developed a CNN-based forecasting framework that extracts spatial information directly from two-dimensional wind fields and combines it with local water-level time series. The approach extended useful forecasts to 72 h and used transfer learning to support application across different stations. However, much of the training database was generated from physics-based numerical storm-surge simulations, making the framework an example of a data-driven surrogate built partly on process-model output rather than an entirely observation-independent alternative. Its computational efficiency is valuable for warning applications, while evaluation on independent observed storms and storm regimes remains important for establishing operational generalization.

Atmospheric dust transport provides another distinct spatiotemporal application. Yarmohamadi et al. (2025) compared CNN-LSTM and convolutional LSTM (ConvLSTM) architectures for dust-pathway prediction at 6-, 12-, 18-, and 24-h lead times using Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) dust and meteorological variables. RF feature importance was used to select contextual predictors such as wind, humidity, and temperature. Incorporating this contextual information improved model performance, and the CNN-LSTM showed higher agreement than the ConvLSTM across the forecast horizons; at 24 h, for example, the reported Kappa coefficients were 0.769 and 0.5465, respectively. The study also illustrates why overall accuracy alone is insufficient: class-sensitive measures such as precision, recall, F1-score, and Kappa provide additional information about the quality of pathway detection. Table 6 summarizes the primary evidence across the distinct storm-related hazard applications reviewed here.

Taken together, these studies show that the main methodological opportunity across storm-related hazards is not a single preferred algorithm but the integration of data structures suited to each physical task. Multimodal sensor fusion is particularly valuable for convective nowcasting, tree-based regression can support rapid reconstruction of cyclone structure, simulation-trained CNN surrogates can reduce the computational burden of storm-surge forecasting, and hybrid convolutional-recurrent models can represent dust transport across space and time. At the same time, credible operational use requires validation that respects storm-level and temporal independence, evaluation across different storm intensities and regions, class-sensitive metrics for rare events, and explicit testing of transfer from reanalysis or simulation-derived training data to real observational conditions.

8.1 Comparative Suitability of Model Families

No model family is universally superior. Tree ensembles and boosting were commonly effective in the reviewed tabular susceptibility, occurrence, and risk-mapping applications, especially when predictors mix scales and distributions. SVMs remain useful for moderate-sized nonlinear datasets, while transductive variants can additionally exploit unlabeled samples in data-scarce settings; however, scaling, kernel choice, and hyperparameter tuning remain important. MLPs capture nonlinear tabular relationships but offer less structural guidance than architectures designed for imagery or sequences. CNNs exploit local spatial structure in satellite images, wind fields, inundation masks, and spread grids; LSTMs and transformers target temporal dependence; and graph models represent irregular spatial connectivity. Hybrid models can combine process simulations, signal decomposition, optimization, or multiple architectures, but every added component expands the tuning space and the risk of optimistic evaluation. A cross-method comparison of the principal ML families, their typical applications, strengths, computational characteristics, and generalization concerns is provided in Table 7.

Figure 3 integrates these cross-method findings into a cross-hazard modeling pipeline linking data and targets, model families, credible evaluation, and decision use.

8.2 Validation Quality, Data Leakage and Rare Events

Random train-test splits are often inadequate for climate and geospatial data because neighboring pixels, stations, catchments, or adjacent time steps share information. A model can therefore interpolate among correlated samples while appearing to generalize. Spatially blocked cross-validation estimates performance in withheld areas; leave-one-region or leave-one-catchment-out designs test geographic transfer; temporal blocking and rolling-origin evaluation preserve chronology; and out-of-period tests assess performance after a regime shift. External validation on a separately collected dataset provides stronger evidence than repeated partitions of the same inventory. The reviewed studies illustrate this distinction: cross-catchment evaluation in rainfall-runoff modeling revealed substantial basin-to-basin variability (Kratzert et al., 2019), while regional testing of FirePred showed that wildfire-spread performance changed outside the primary training region (Marjani et al., 2023).

All data-dependent operations must be fitted inside each training fold.

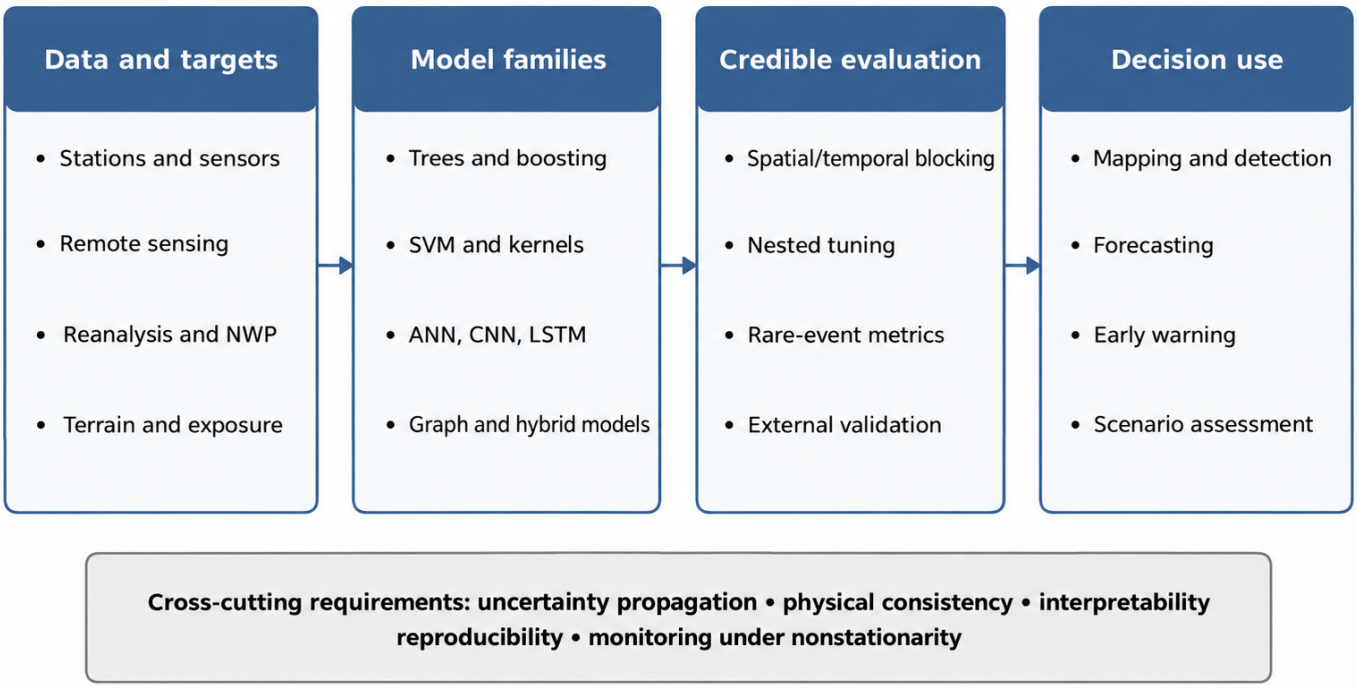


Figure 3 Cross-Hazard Modeling Pipeline. Credible Evaluation and Cross-Cutting Safeguards Connect Data-Driven Performance to Decision Use. NWP, Numerical Weather Prediction; SVM, Support Vector Machine; ANN, Artificial Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory.

Table 8 Evaluation Threats and Stronger Alternatives

Threat	Why performance can be optimistic	Preferred evaluation
Spatial autocorrelation	Nearby training and test samples share terrain, climate, or event labels	Spatial blocks; leave-region/catchment-out; external geography
Temporal dependence	Random splitting places adjacent observations in both sets	Blocked time split; rolling origin; out-of-period test
Preprocessing leakage	Scaling, imputation, feature selection, or decomposition uses full-data information.	Fit the complete pipeline within each training fold.
Resampling leakage	SMOTE or undersampling before splitting duplicates neighborhood information	Resample training folds only; test on natural prevalence.
Hyperparameter overfitting	Repeated tuning on the test set inflates final performance	Nested cross-validation or untouched final test set
Class imbalance	Accuracy and ROC-AUC can obscure rare-event errors	Precision-recall, class-specific recall, F1, calibration, event metrics
Domain shift	Climate, sensors, land use, or event definitions differ at deployment	Out-of-region/out-of-period testing; adaptation and monitoring
Sampling or event-inventory bias	Event locations may overrepresent well-observed or accessible areas, while non-event or pseudo-absence samples may not represent the full environmental domain.	Report sampling strategy; test sensitivity to event/non-event construction; use spatially separated or independent inventories where available

Scaling before splitting leaks test-distribution information. Selecting predictors on the full dataset biases feature evaluation. Applying SMOTE or other resampling before splitting can place synthetic relatives of test cases in training. Hyperparameters selected using the nominal test set convert it into validation data. Nested cross-validation or a separate final test set is therefore necessary when extensive tuning is performed. These are methodological risks; they should not be attributed to an individual paper unless its published workflow establishes that leakage occurred. Accordingly, the present review treats these issues as evaluation risks unless the published methodology provides sufficient information to determine how preprocessing, resampling, or tuning was separated from final testing.

Rare hazards require metrics tied to decisions. Accuracy can be dominated by non-events. Under severe class imbalance, ROC-AUC should be interpreted alongside class-specific measures and precision-recall behavior rather than used as the sole indicator of performance. Precision, recall, F1, probability calibration, event-based hit and false-alarm rates, lead-time-dependent error, and spatial overlap or boundary metrics should be selected according to the task. Thresholds should be chosen without using the final test data and accompanied by uncertainty intervals. This issue is evident in Choudary et al. (2025), where a model achieved very high overall accuracy while precision, recall, and F1 for the heatwave class were zero, demonstrating that aggregate accuracy alone can conceal failure on the rare event of interest. Conversely, Mecikalski et al. (2021) reported several event-sensitive measures, including CSI, POD, false-alarm-related metrics, and skill scores, providing a more informative evaluation of severe-storm warnings than accuracy alone. Table 8 summarizes the main evaluation threats identified across the reviewed applications and the corresponding stronger evaluation alternatives.

8.3 Uncertainty, Physical Consistency and Interpretability

Uncertainty enters at several levels. Observations contain measurement error, missingness, and inconsistent event labels; remote-sensing retrievals add cloud, resolution, and algorithm uncertainty; reanalysis and numerical forecasts inherit model and assimilation error; and future assessments add climate-model, downscaling, emissions-scenario, land-use, and socioeconomic uncertainty. ML contributes predictive uncertainty from finite samples, model structure, optimization, and distribution shift.

Uncertainty analysis should reflect the structure of the application. For spatial mapping, this includes sensitivity to inventories, predictors, and geographic transfer; for forecasting, uncertainty should be examined across lead times and out-of-period conditions; and for scenario-based studies,

results should be interpreted across the climate models, scenarios, and socioeconomic assumptions used. Where models generate probabilities or prediction intervals, their calibration and coverage should be evaluated rather than reporting point estimates alone.

Physical credibility can be strengthened by coupling ML with process-model information, using ML for residual correction or simulation-based surrogates, and checking predictions against physically plausible ranges and known behavior. A fast surrogate may complement a process model without reproducing every governing mechanism. XAI methods such as SHAP, PDP, Expected Gradients, and interaction values describe the fitted model and can reveal relationships consistent with domain knowledge. Consistency is useful evidence, but explanation is not causation and does not enforce a physical law.

9. FUTURE RESEARCH DIRECTIONS

Future research should prioritize evidence that models remain reliable across changes in location, time, and observation system. Multi-region evaluation should reserve entire catchments, cities, ecosystems, or event periods for independent testing, while rolling-origin and out-of-period evaluation can reveal performance degradation under nonstationary conditions. Transfer learning and domain-adaptation approaches should be compared with local recalibration and appropriate process-based or statistical baselines, with explicit reporting of both successful and unsuccessful transfer.

Uncertainty should be incorporated explicitly into model evaluation and future-risk assessment rather than represented only by point estimates. Probabilistic forecasts, calibrated prediction intervals, ensemble approaches, and multi-model or multi-scenario assessments can help characterize uncertainty relevant to operational thresholds and long-term projections. Rare and unprecedented extremes also require event-specific evaluation and explicit examination of model behavior outside the range represented by historical training data.

Multimodal data integration remains an important opportunity because climate hazards are observed through complementary station, radar, satellite, reanalysis, forecast-model, terrain, ecological, infrastructure, and socioeconomic information. Future studies should distinguish improvements arising from genuinely complementary information from those resulting primarily from increased resolution or inadvertent information leakage. Fusion strategies should also account for missing observations, differences in temporal resolution, sensor drift, and uncertainty in individual data sources.

Reproducibility and operational evaluation should receive greater emphasis. Open benchmark datasets, documented and versioned preprocessing, public code where possible, fixed evaluation splits, and complete hyperparameter reporting would make comparisons across studies more reliable. Operational studies should additionally report computational cost, latency, update frequency, calibration drift, failure modes, and the way model outputs are incorporated into human decision processes. Operational value should therefore be assessed jointly in terms of predictive performance, computational efficiency, interpretability, robustness, and decision latency.

Future process-guided and physics-informed approaches should specify which physical information or constraints are incorporated and demonstrate their value relative to appropriate data-driven and process-based baselines. Similarly, interpretability methods should be evaluated for stability and domain relevance, while causal claims require evidence beyond feature attribution alone. ML should remain complementary to process-based and statistical approaches: it can support observation fusion, nonlinear pattern recognition, surrogate modeling, and systematic-error correction, while process-based models provide mechanistic structure and important support for scenario interpretation.

10. CONCLUSION

This review examined machine-learning applications across five consistent climate-hazard categories: floods and stormwater, droughts, temperature extremes, wildfires, and storm-related hazards. The evidence shows broad but strongly task-dependent value for ML. Tree-based ensembles and boosting methods have been widely applied to tabular susceptibility, occurrence, and risk-mapping problems, while SVMs remain useful for moderate-sized nonlinear datasets and convolutional, recurrent, transformer, graph-based, and hybrid architectures extend modeling to imagery and complex spatiotemporal processes. Direct within-study comparisons sometimes demonstrate advantages for particular models under shared data and evaluation conditions; however, differences in prediction targets, spatial and temporal scales, class distributions, forecast horizons, datasets, and validation protocols prevent a defensible universal ranking of model families.

Recurring scientific gaps concern validation, generalization, uncertainty, and operational deployment. Reported performance can be sensitive to spatial and temporal dependence, event rarity, sampling design, preprocessing and tuning procedures, and distribution shifts between model-development and deployment conditions. Geographic transfer and external validation remain limited in many applications, while scenario-based assessments and multimodal models introduce additional uncertainty from observational products, climate projections, and heterogeneous data sources. Explainability methods can improve transparency regarding fitted-model behavior, but their attributions should not be interpreted as direct evidence of physical causality.

Accordingly, ML should be viewed as a complementary component of climate-hazard modeling rather than a universal replacement for statistical or process-based approaches. Its practical value will depend on representative data, rigorous and independent validation, transparent uncertainty characterization, reproducible modeling workflows, consistency with relevant physical understanding, and demonstrated reliability under the regions, lead times, event classes, and operational conditions for which the models are intended.

REFERENCES

Ali, Z., Hussain, I., Faisal, M., Nazir, H. M., Hussain, T., Shad, M. Y., Shoukry, A. M., & Gani, S. H. (2017). Forecasting drought using a multilayer perceptron artificial neural network model. *Advances in Meteorology, 2017*(1), 5681308. https://doi.org/10.1155/2017/5681308

Alshahrani, M. A., Laiq, M., Noor-ul-Amin, M., Yasmeen, U., & Nabi, M. (2024). A support vector machine based drought index for regional drought analysis. *Scientific Reports, 14*(1), 9849. https://doi.org/10.1038/s41598-024-60616-3

Amiri, A., Soltani, K., Ebtehaj, I., & Bonakdari, H. (2024). A novel machine learning tool for current and future flood susceptibility mapping by integrating remote sensing and geographic information systems. *Journal of Hydrology, 632*, 130936. https://doi.org/10.1016/j.jhydrol.2024.130936

Bhoopathi, S., Kumar, N., Somesh, & Pal, M. (2024). Evaluating the performances of SVR and XGBoost for short-range forecasting of heatwaves across different temperature zones of India. *Applied Computing and Geosciences, 24*, 100204. https://doi.org/10.1016/j.acags.2024.100204

Bhowmik, R. T., Jung, Y. S., Aguilera, J. A., Prunicki, M., & Nadeau, K. (2023). A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework. *Journal of Environmental Management, 341*, 117908. https://doi.org/10.1016/j.jenvman.2023.117908

Buschow, S., Keller, J., & Wahl, S. (2024). Explaining heatwaves with machine learning. *Quarterly Journal of the Royal Meteorological Society, 150*(760), 1207–1221. https://doi.org/10.1002/qj.4642

Caiado, J., & Marques, M. (2025). Predicting wildfire occurrences in Portugal using machine learning classification models. *Ecological Informatics, 92*, 103455. https://doi.org/10.1016/j.ecoinf.2025.103455

Cheraghi Ghalehsari, A., Habibnejad Roshan, M., & Roshun, S. H. (2020). Flood susceptibility mapping using a support vector machine models (SVM) and geographic information system (GIS). *Journal of Natural Environmental Hazards, 9*(25), 61–80. https://doi.org/10.22111/jneh.2020.31018.1547

Choudary, V. R., Johnvictor, A. C., & Sankar N. P. (2025). Comparative analysis of machine learning approaches for heatwave event prediction in India.

Scientific Reports, 15, 22431. https://doi.org/10.1038/s41598-025-04634-9

Dai, H., Xiong, L., Ma, Q., & Duan, Z. (2025). Deep learning model for drought prediction based on large-scale spatial causal network in the Yangtze River Basin. *Journal of Hydrology, 654*, 132808. https://doi.org/10.1016/j.jhydrol.2025.132808

Duan, Y., Bo, Y., Yao, X., Chen, G., Liu, K., Wang, S., Yang, B., & Li, X. (2025). A deep learning framework for long-term soil moisture-based drought assessment across the major basins in China. *Remote Sensing, 17*(6), 1000. https://doi.org/10.3390/rs17061000

Enung, Kasyanto, H., Sari, R. R., & Lubis, M. F. (2023, June). Application of multilayer perceptron (MLP) method for streamflow forecasting (case study: Upper Citarum River, Indonesia). In *IOP Conference Series: Earth and Environmental Science, 1203*(1), 012032. IOP Publishing. https://doi.org/10.1088/1755-1315/1203/1/012032

Eslaminezhad, S. A., Eftekhari, M., Mahmoodizadeh, S., Akbari, M., & Haji Elyasi, A. (2021). Evaluation of tree-based artificial intelligence models to predict flood risk using GIS. *Iran-Water Resources Research, 17*(2), 174–189.

Frassoni, A., Reynolds, C., Wedi, N., Bouallègue, Z. B., Caltabiano, A. C. V., Casati, B., Christophersen, J. A., Coelho, C. A. S., De Falco, C., Doyle, J. D., Fernandes, L. G., Forbes, R., Janiga, M. A., Klocke, D., Magnusson, L., McTaggart-Cowan, R., Pakdaman, M., Rushley, S. S., Verhoef, A.,... Zängl, G. (2023). Systematic errors in weather and climate models: Challenges and opportunities in complex coupled modeling systems. *Bulletin of the American Meteorological Society, 104*(9), E1687–E1693. https://doi.org/10.1175/BAMS-D-23-0102.1

Gkirmpas, P., Tsegas, G., Boehnke, D., Vlachokostas, C., & Moussiopoulos, N. (2025). Temperature prediction at street scale during a heat wave using random forest. *Atmosphere, 16*(7), 877. https://doi.org/10.3390/atmos16070877

Goumghar, L., Hajaj, S., Haida, S., Kili, M., Mridekh, A., Khandouch, Y.,... & El Mansouri, B. (2025). Analysis of baseline and novel boosting models for flood-prone prediction and explainability: Case from the Upper Drâa Basin (Morocco). *Earth, 6*(3), 69. https://doi.org/10.3390/earth6030069

He, Q., Zhu, Z., Zhao, D., Song, W., & Huang, D. (2024). An interpretable deep learning approach for detecting marine heatwaves patterns. *Applied Sciences, 14*(2), 601. https://doi.org/10.3390/app14020601

Huang, C., She, D., Liu, X., Zhang, Y., Wu, Y., Shi, X., & Wei, Y. (2025). Predicting future meteorological drought risk in mainland China using random forest model. *Journal of Hydrology: Regional Studies, 61*, 102633. https://doi.org/10.1016/j.ejrh.2025.102633

Ishak, S. I., Sitanggang, I. S., & Widodo, T. (2024, June). Comparison of K-Means and K-Medoids algorithms for clustering of potential flood-prone areas in Bengkulu Province. In *IOP Conference Series: Earth and Environmental Science, 1359*(1), 012017. IOP Publishing. https://doi.org/10.1088/1755-1315/1359/1/012017

Jiang, W., Qiao, Y., Su, G., Li, X., Meng, Q., Wu, H., Quan, W., Wang, J., & Wang, F. (2023). WFNet: A hierarchical convolutional neural network for wildfire spread prediction. *Environmental Modelling & Software, 170*, 105841. https://doi.org/10.1016/j.envsoft.2023.105841

Jiang, H., Ren, T., Huang, R., & Yu, K. (2025). Marine heatwaves and cold spells in global coral reef regions (1982–2070): Characteristics, drivers, and impacts. *Remote Sensing, 17*(16), 2881. https://doi.org/10.3390/rs17162881

Kan, J. C., Vieira Passos, M., Destouni, G., Barquet, K., Ferreira, C. S. S., & Kalantari, Z. (2025). Seasonal heatwave forecasting with explainable machine learning and remote sensing data. *Stochastic Environmental Research and Risk Assessment, 39*, 3333–3352. https://doi.org/10.1007/s00477-025-03020-1

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences, 23*, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Langhammer, J. (2023). Flood simulations using a sensor network and support vector machine model. *Water, 15*(11), 2004. https://doi.org/10.3390/w15112004

Lee, J. H., Kim, J. S., Jang, H. W., & Lee, J. C. (2013). Drought forecasting using the multilayer perceptron (MLP) artificial neural network model. *Journal of Korea Water Resources Association, 46*(12), 1249–1263. https://doi.org/10.3741/JKWRA.2013.46.12.1249

Lee, S., Kim, J. C., Jung, H. S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk, 8*(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

Li, H., Huang, J., Zhang, X., Meng, Z., Fan, Y., Wu, X., Wang, L., Hu, L., & Zhang, J. (2025). Flash flood risk classification using GIS-based fractional order k-means clustering method. *Fractal and Fractional, 9*(9), 586. https://doi.org/10.3390/fractalfract9090586

Li, M., Yao, Y., Feng, Z., & Ou, M. (2025). Hydrological drought prediction and its influencing features analysis based on a machine learning model. *Natural Hazards and Earth System Sciences, 25*(11), 4299–4316. https://doi.org/10.5194/nhess-25-4299-2025

Li, J., Zheng, A., Guo, W., Bandyopadhyay, N., Zhang, Y., & Wang, Q. (2023). Urban flood risk assessment based on DBSCAN and K-means clustering algorithm. *Geomatics, Natural Hazards and Risk, 14*(1), 2250527. https://doi.org/10.1080/19475705.2023.2250527

Liu, J., Wang, J., Xiong, J., Cheng, W., Li, Y., Cao, Y., He, Y., Duan, Y., He, W., & Yang, G. (2022). Assessment of flood susceptibility mapping using support vector machine, logistic regression and their ensemble techniques in the Belt and Road region. *Geocarto International, 37*(25), 9817–9846. https://doi.org/10.1080/10106049.2022.2025918

Lopez-Gomez, I., McGovern, A., Agrawal, S., & Hickey, J. (2023). Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems, 2*(1), e220035. https://doi.org/10.1175/AIES-D-22-0035.1

Ma, T., Wang, G., Guo, R., Chen, L., & Ma, J. (2024). Forest fire susceptibility assessment under small sample scenario: A semi-supervised learning approach using transductive support vector machine. *Journal of Environmental Management, 359*, 120966. https://doi.org/10.1016/j.jenvman.2024.120966

Marina, C. M., Pérez-Aracil, J., McAdam, R., Lorente-Ramos, E., Luther, N., Zorita, E., Scoccimarro, E., Luterbacher, J., Xoplaki, E., & Salcedo-Sanz, S. (2026). Detection and attribution of heat waves with the Multivariate Autoencoder Flow-Analogue Method (MvAE-AM). *Atmospheric Research, 328*, 108409. https://doi.org/10.1016/j.atmosres.2025.108409

Marjani, M., Ahmadi, S. A., & Mahdianpari, M. (2023). FirePred: A hybrid multi-temporal convolutional neural network model for wildfire spread prediction. *Ecological Informatics, 78*, 102282. https://doi.org/10.1016/j.ecoinf.2023.102282

Mecikalski, J. R., Sandmæl, T. N., Murillo, E. M., Homeyer, C. R., Bedka, K. M., Apke, J. M., & Jewett, C. P. (2021). A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar-GOES satellite-lightning observations. *Monthly Weather Review, 149*(6), 1725–1746. https://doi.org/10.1175/MWR-D-19-0274.1

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536. https://doi.org/10.3390/w10111536

Nemni, E., Bullock, J., Belabbes, S., & Bromley, L. (2020). Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. *Remote Sensing, 12*(16), 2532. https://doi.org/10.3390/rs12162532

Pakdaman, M., Babaeian, I., & Bouwer, L. M. (2022). Improved monthly and seasonal multi-model ensemble precipitation forecasts in southwest Asia using machine learning algorithms. *Water, 14*(17), 2632. https://doi.org/10.3390/w14172632

Peterson, T. C., Heim, R. R., Jr., Hirsch, R. M., Kaiser, D. P., Brooks, H., Diffenbaugh, N. S., Dole, R. M., Giovannettone, J. P., Guirguis, K., Karl, T. R., Katz, R. W., Kunkel, K. E., Lettenmaier, D. P., McCabe, G. J., Paciorek, C. J., Ryberg, K. R., Schubert, S., Silva, V. B. S., Stewart, B. C.,... Wuebbles, D. (2013). Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: State of knowledge. *Bulletin of the American Meteorological Society, 94*(6),

821–834. https://doi.org/10.1175/BAMS-D-12-00066.1

Ramachandran, R., Arthi, D., Sudha, D., Anil, A. B., Babu, A., & Haripriya, M. D. (2024, December). Harnessing QGIS and random forest analysis for precision storm drain reinforcement. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICERCS63125.2024.10895146

Razavi-Termeh, S. V., Pourzangbar, A., Sadeghi-Niaraki, A., Franca, M. J., & Choi, S. M. (2025). Metaheuristic-driven enhancement of categorical boosting algorithm for flood-prone areas mapping. *International Journal of Applied Earth Observation and Geoinformation, 136*, 104357. https://doi.org/10.1016/j.jag.2025.104357

Shada, B., Chithra, N. R., & Thampi, S. G. (2022). Hourly flood forecasting using hybrid wavelet-SVM. *Journal of Soft Computing in Civil Engineering, 6*(2), 1–20. https://doi.org/10.22115/scce.2022.317761.1383

Shan, B., Verhoest, N. E., & De Baets, B. (2024). Identification of compound drought and heatwave events on a daily scale and across four seasons. *Hydrology and Earth System Sciences, 28*(9), 2065–2080. https://doi.org/10.5194/hess-28-2065-2024

Singha, C., Swain, K. C., Moghimi, A., Foroughnia, F., & Swain, S. K. (2024). Integrating geospatial, remote sensing, and machine learning for climate-induced forest fire susceptibility mapping in Similipal Tiger Reserve, India. *Forest Ecology and Management, 555*, 121729. https://doi.org/10.1016/j.foreco.2024.121729

Tanriverdi, I., & Batmaz, İ. (2025). AI-driven US drought prediction using machine learning and deep learning. *Climate Dynamics, 63*(6), 249. https://doi.org/10.1007/s00382-025-07720-w

Tran, T. T. K., Janizadeh, S., Bateni, S. M., Jun, C., Kim, D., Trauernicht, C., Rezaie, F., Giambelluca, T. W., & Panahi, M. (2024). Improving the prediction of wildfire susceptibility on Hawai'i Island, Hawai'i, using explainable hybrid machine learning models. *Journal of Environmental Management, 351*, 119724. https://doi.org/10.1016/j.jenvman.2023.119724

Wahba, M., Essam, R., El-Rawy, M., Al-Arifi, N., Abdalla, F., & Elsadek, W. M. (2024). Forecasting of flash flood susceptibility mapping using random forest regression model and geographic information systems. *Heliyon, 10*(13), e33982. https://doi.org/10.1016/j.heliyon.2024.e33982

Wang, Y., Fang, Z., Hong, H., Costache, R., & Tang, X. (2021). Flood susceptibility mapping by integrating frequency ratio and index of entropy with multilayer perceptron and classification and regression tree. *Journal of Environmental Management, 289*, 112449. https://doi.org/10.1016/j.jenvman.2021.112449

Wang, Y., Fang, Z., Hong, H., & Peng, L. (2020). Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology, 582*, 124482. https://doi.org/10.1016/j.jhydrol.2019.124482

Wang, H., Meng, Y., Wang, H., Wu, Z., & Guan, X. (2023). The application of integrating comprehensive evaluation and clustering algorithms weighted by maximal information coefficient for urban flood susceptibility. *Journal of Environmental Management, 344*, 118846. https://doi.org/10.1016/j.jenvman.2023.118846

Wang, Q., Zhao, L., Wang, M., Wu, J., Zhou, W., Zhang, Q., & Deng, M. (2022). A random forest model for drought: Monitoring and validation for grassland drought based on multi-source remote sensing data. *Remote Sensing, 14*(19), 4981. https://doi.org/10.3390/rs14194981

Welandawe, S., Priyadarshana, Y. H. P. P., Senanayake, N., & Silva, E. N. S. (2025). Machine learning techniques for marine heatwave prediction: A comprehensive review. *Intelligent Marine Technology and Systems, 3*, 28. https://doi.org/10.1007/s44295-025-00076-1

Xie, W., Xu, G., Zhang, H., & Dong, C. (2023). Developing a deep learning-based storm surge forecasting model. *Ocean Modelling, 182*, 102179. https://doi.org/10.1016/j.ocemod.2023.102179

Yarmohamadi, M., Alesheikh, A. A., & Sharif, M. (2025). Using hybrid deep learning models to predict dust storm pathways with enhanced accuracy. *Climate, 13*(1), 16. https://doi.org/10.3390/cli13010016

Youn, J., Hwang, S., Qian, X., Im, S., & Son, S. (2026). A random forest-based method for reconstructing pressure fields in tropical cyclones from wind data. *Atmospheric Research, 330*, 108588. https://doi.org/10.1016/j.atmosres.2025.108588

Zandler, H., & Samimi, C. (2024). Cooling potential of urban tree species during extreme heat and drought: A thermal remote sensing assessment. *Remote Sensing, 16*(12), 2059. https://doi.org/10.3390/rs16122059