https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

A Basin-Aware Global Framework for Computationally Efficient Surface Water Inundation Prediction

Arik M. Tashie[1], Isaac D. Gerg[1], Evan Koester[1], Carlos D. Hoyos[1], Eduardo Galindo[1], and David J. Farnham[1]

[1]ClimateAi, San Francisco, California, USA
[1]*Correspondence to:* Arik M. Tashie (arik@climate.ai)

Abstract. Predicting surface water inundation at regional to global scales presents a fundamental tension: bespoke local models achieve high accuracy but require proprietary data and are difficult to scale, while globally trained systems offer broad coverage but demand substantial computational infrastructure and may lack flexibility for regional customization. We present the Basin-Aware Global Inundation Modeling framework (BAGIM), which addresses this gap by combining globally available, freely accessible datasets with basin-scale calibration to capture regional hydrological specificity. The BAGIM framework is model-architecture agnostic, and facilitates the prediction of daily snapshots of inundation extent (the binary footprint of surface water on a given day). We evaluate six model architectures across eight geographically diverse basins to test three hypotheses: (1) that hydrologically meaningful feature engineering is more impactful than architectural complexity, (2) that basin-scale training mitigates regional biases in global datasets, and (3) that basin-aware models can generalize to extreme events beyond the training distribution. Our experiments demonstrate that tree-based ensembles (XGBoost, Random Forest) consistently outperform more complex deep learning architectures, achieving median F1 scores of approximately 0.5 against OPERA DSWx-S1 reference snapshots, performance that approaches the inherent uncertainty ceiling imposed by disagreement among remote sensing products themselves in settings with small, shallow, and intermittent water bodies. We find that features commonly assumed essential for operational flood forecasting (i.e., coincident river-basin streamflow, Height Above Nearest Drainage, and elevation) are neither sufficient nor strictly necessary for reliable prediction, with well-engineered meteorological and terrain features achieving comparable performance without explicit streamflow inputs. This challenges a core assumption underlying many current operational flood forecasting systems. Cross-basin transfer experiments reveal limited transferability, reinforcing the importance of basin-aware calibration. Further, models trained exclusively on non-extreme events produce directionally correct predictions for out-of-sample extremes, though with conservative bias (higher precision, lower recall). We suggest that a design philosophy prioritizing feature engineering and regional calibration over architectural complexity enables accessible, globally deployable inundation mapping without sacrificing predictive skill.

1 Introduction

Surface water inundation (including floods, reservoirs, small water bodies, and wetlands) plays a central role in shaping hydrological, ecological, and socio-economic systems worldwide (Pekel et al., 2016; Tharme, 2003; Yamazaki et al., 2011).

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Accurate and timely mapping of inundated versus non-inundated landscapes is essential for disaster response, water resource management, ecosystem monitoring, and climate adaptation planning (e.g., infrastructure siting, insurance risk modeling, agricultural water management) (Sheffield et al., 2012; Sajjad et al., 2023; Inman and Lyons, 2020). While catastrophic floods are a major hazard, the broader challenge of general inundation modeling includes the delineation of reservoirs for water supply, the detection of ephemeral and small water bodies critical for biodiversity, and the assessment of drought impacts on surface water availability (Wang et al., 2022; Ferreira et al., 2018; Trochim et al., 2016).

Traditionally, the simulation and prediction of inundation extent have relied on physics-based hydrodynamic models, which solve simplified forms of the shallow-water equations to simulate the propagation of flood waves through a landscape at sub-daily time steps, yielding time-varying fields of water depth and velocity. These models require detailed topographic, hydraulic, and boundary condition data, and are often computationally intensive and difficult to calibrate at large scales (Neal et al., 2012; Teng et al., 2017). The proliferation of remote sensing data, particularly from the Sentinel and Landsat missions, has enabled the development of global surface water products, such as NASA's OPERA DSWx suite, that provide near-real-time, high-resolution water masks derived from both optical and synthetic aperture radar (SAR) data (Laboratory, 2023). These advances spurred a parallel rise in empirical and machine learning (ML) approaches for surface water mapping that leverage rich remote sensing and ancillary geospatial datasets to infer inundation patterns across diverse landscapes and hydrological regimes (Rosser et al., 2017; Yang et al., 2018; Acharya et al., 2019).

It is useful to distinguish two related but distinct modeling objectives that are sometimes conflated under the umbrella of "flood modeling". *Hydrodynamic flood models* simulate the continuous propagation of a flood wave through a river network and its floodplain, producing time-varying estimates of water depth, velocity, and extent within a single event. *Inundation extent models* ask a complementary question: which areas are inundated at a given moment? These models produce binary or probabilistic footprints rather than depth hydrographs, and their reference data are typically remote sensing snapshots acquired at discrete times. The empirical and ML approaches cited above tend to fall into the second category and are well suited to applications where the daily spatial footprint of surface water is more actionable than the hour-by-hour rise and fall of water elevation within a single event. For example, disaster response triage, reservoir and small water body monitoring, exposure assessment, and climate adaptation planning all require knowing where water is on a given day, often across many basins and many days at once, rather than resolving the sub-daily hydrodynamics of any individual flood pulse. The work presented here sits within this second category.

Recent years have witnessed the operationalization of global flood forecasting systems such as Google Flood Hub and GloFAS, which integrate streamflow forecasts, remote sensing, and digital elevation models (DEMs) to provide flood alerts and water extent maps at continental to global scales (Zsoter et al., 2020; Arad et al., 2022). These systems represent a significant advance for large-scale flood monitoring and early warning. However, global systems by design prioritize consistency and scalability, often linking streamflow predictions to inundation extent via static flood masks or topographic thresholds. Because the accuracy of remote sensing-derived water masks is subject to substantial uncertainty, especially for shallow inundation, vegetated or urban environments, and small water features (Martinis et al., 2022; Uday et al., 2025; Cao et al., 2024), this one-to-one relationship can impose a ceiling on the achievable skill of any data-driven model. Therefore, complementary approaches

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

are needed to address use cases where flexibility, regional calibration, or computational tractability are paramount. This need is pronounced in regions dominated by small, fragmented, or ephemeral water bodies, or where local hydrogeomorphic and land use variability is high.

In response to these challenges, the hydrological community has begun to explore hybrid and physics-informed ML frameworks that seek to combine the strengths of process-based understanding (Cuomo et al., 2022; Beven and Binley, 1992; Tashie et al., 2022) with the flexibility and scalability of empirical models (Nearing et al., 2021; Sit et al., 2020; Kabir et al., 2020). Ensemble and transfer learning methods, as well as uncertainty quantification strategies, are increasingly recognized as essential for robust, generalizable, and operationally useful inundation mapping (Sharma and Saharia, 2025; Mangukiya et al., 2024; Zhao et al., 2021). Yet, critical questions remain regarding the optimal balance between model complexity and hydrological realism, the impact of regional data biases and training strategies, and the ability of empirical models to generalize to out-of-sample events and locations (Gauch et al., 2021; Frame et al., 2020; Zhang et al., 2024).

This spectrum of approaches reveals a practical gap: bespoke local models can achieve high accuracy but require proprietary or local data and are difficult to scale; while globally trained systems offer broad coverage but demand substantial computational infrastructure and may be less flexible for rapid adaptation or regional customization. We propose that a globally applicable framework that trains and validates at the basin scale could offer both regional sensitivity and global operational tractability.

In this work, we present the Basin-Aware Global Inundation Modeling framework (BAGIM): an empirical ML approach designed to be globally applicable yet regionally calibrated, reliant solely on freely available static geophysical and dynamic meteorological data. By training models at the basin scale within a unified methodological framework, we aim to capture regional specificity while maintaining the scalability and reproducibility of a globally consistent approach. Our framework is intentionally broad, targeting not only floods but also the delineation of reservoirs, small water bodies, and water supply status across a wide range of landscapes (Figure 1). We explicitly recognize that both input features (e.g., DEMs, land use and land cover or "LULC", soils, infrastructure) and ground truth targets (e.g., inundation masks from OPERA DSWx-S1) are subject to substantial uncertainty, particularly in data-scarce or hydrologically complex regions (Venter et al., 2022; Wechsler, 2003; Hengl et al., 2017). Rather than optimizing solely for traditional accuracy metrics, we emphasize the importance of hydrologically realistic inference, physical plausibility, and robust generalization across diverse inundation contexts.

Specifically, we test the following hypotheses:

- **H1 – Feature engineering over model complexity**: Within the BAGIM framework, hydrologically meaningful feature engineering yields more robust inundation predictions than increased architectural complexity alone, thus enabling computationally efficient deployment without specialized infrastructure.
- **H2 – Basin-scale training mitigates regional bias**: Regional biases in global geospatial datasets propagate into globally uniform inundation models, but can be mitigated by training at the basin scale within a unified methodological framework, producing consistent performance across diverse hydroclimatic regimes.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

– **H3 – Generalization to out-of-sample events**: A BAGIM model can generalize to predict out-of-sample inundation both spatially (in held-out watersheds) and temporally (for extreme events beyond the training distribution) when supplied with sufficiently rich hydrometeorological and geomorphic context.

To address these hypotheses, we develop and evaluate a suite of ML models, including tree-based ensembles, deep neural networks, and hybrid architectures across multiple basins and inundation regimes. Then we systematically assess the impact of feature engineering, model complexity, transfer learning, and ensemble aggregation on predictive skill, robustness to label noise, and generalization to extreme events.

We selected eight study basins from the STURM-Flood database (Notarangelo et al., 2025) to satisfy two criteria: (1) availability of coincident Sentinel-1 and Sentinel-2 coverage during flood events, enabling analysis of model performance across alternative validation products, and (2) representation of diverse hydroclimates and flood regimes, including both pluvial and fluvial flooding mechanisms (Figure 1). The eight study basins span three continents and include sites in Togo (1), Ukraine (2), UK (3), Madagascar (4), Uganda (5), Greece (6), Pakistan (7), and Spain (8). Locations 1, 3, and 6 (Togo, UK, Greece) correspond to pluvial flash-flood events; locations 2, 4, 7, and 8 (Ukraine, Madagascar, Pakistan, Spain) correspond to fluvial (riverine) flood events; and location 5 (Uganda) is a lakeside flood along the shoreline of Lake Victoria.

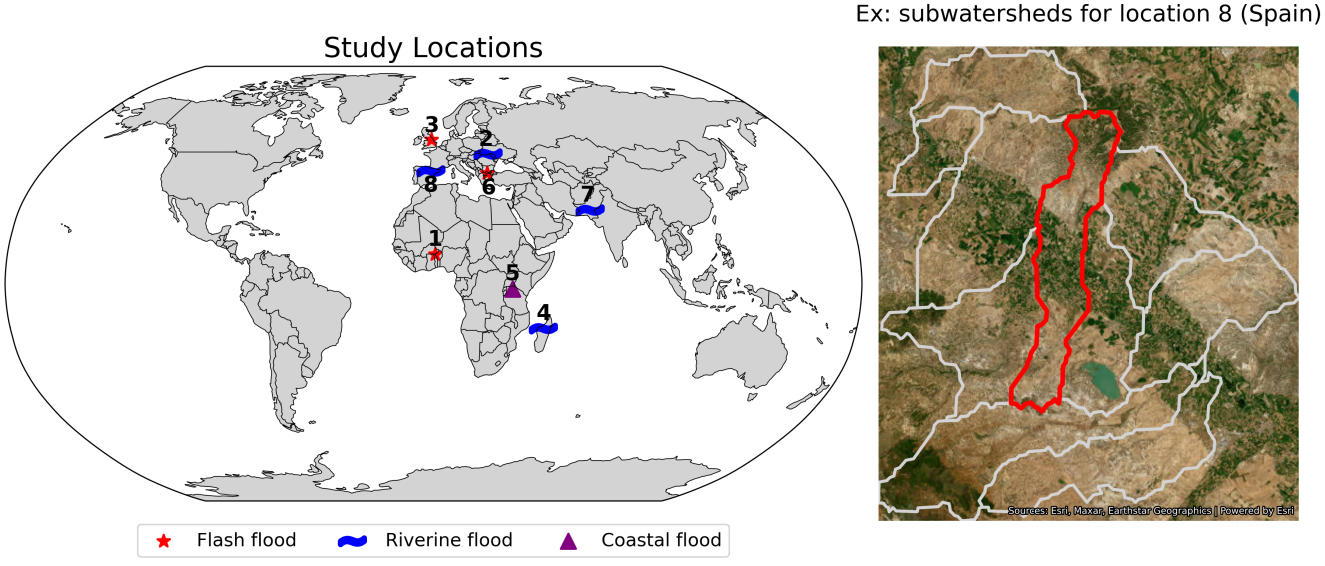

Figure 1. (Left) Study locations and associated flood type in STURM-Flood database. (Right) Example target subwatershed (red) and proximal subwatersheds used in training and validation (grey).

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

2 Methods

Section 2.1 provides an overview of the framework's inputs, outputs, and intended use cases. Section 2.2 details the data sources (summarized in Table 1) used to construct input features. Section 2.3 describes our streamflow modeling approach. Section 2.4 presents the six model architectures evaluated and the common training framework.

2.1 Framework Overview

BAGIM predicts surface water inundation extent at 30 m spatial resolution for any basin where globally available data products are accessible. The framework addresses a fundamental question in hydrological monitoring: given current meteorological conditions and landscape characteristics, which areas are likely to be inundated?

Inputs: BAGIM combines two categories of inputs: (1) *dynamic forcings* consisting of daily meteorological variables (including temperature, precipitation, solar radiation, and humidity) and modeled streamflow, representing current and antecedent hydrometeorological conditions; and (2) *static covariates* describing terrain, soil texture, and land use and land cover at each pixel location. While BAGIM is designed to function with globally available datasets alone, the modular architecture readily accommodates local or bespoke data sources (e.g., higher-resolution DEMs, gauged streamflow, local land cover products) where such data may improve prediction accuracy, though that analysis is outside the scope of this manuscript.

Outputs: The model produces a binary inundation mask at 30 m resolution, classifying each pixel as either inundated or non-inundated for a given date. Probabilistic outputs can also be extracted prior to threshold application.

Goals: BAGIM aims to provide operationally useful inundation predictions that are (1) globally applicable without requiring proprietary or locally-specific data, (2) computationally tractable without specialized GPU infrastructure, and (3) flexible enough to accommodate regional calibration, local data integration, and algorithmic updates without requiring global retraining.

Operational use: To deploy BAGIM for a new basin, users delineate the target watershed, assemble the globally available input datasets (supplemented with local data where beneficial), and train a basin-specific model using available remote sensing-derived inundation observations. Once trained, the model can generate inundation predictions for any date given corresponding meteorological inputs.

2.2 Data Sources and Preprocessing

BAGIM relies exclusively on globally available datasets for both dynamic forcings and static covariates (Table 1). This design enables regional specificity through basin-scale calibration while maintaining methodological consistency across diverse hydroclimatic settings. We deliberately use multiple DEM-derived products because different global elevation datasets exhibit regionally varying biases due to differences in sensor characteristics, processing algorithms, and acquisition dates (Yamazaki et al., 2020); incorporating both allows models to learn which elevation representation is more reliable in different contexts. These datasets are combined with streamflow estimates from a long short-term memory network with feature-wise linear modulation (LSTM-FiLM); the architecture and training of the streamflow model are described in Section 2.3.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Data Source	Description	Justification
ERA5 (Hersbach et al., 2020)	Daily meteorological reanalysis (temperature, precipitation, radiation)	Global coverage; consistent temporal resolution; captures antecedent conditions
NASADEM (Crippen et al., 2016)	30 m digital elevation model	Terrain metrics and elevation anomalies for topographic position
COPDEM GLO-30 HAND (Nobre et al., 2011)	30 m Height Above Nearest Drainage	First-order proxy for inundation susceptibility
SoilGrids 2.0 (Poggio et al., 2021)	Soil texture (clay, sand, silt) at multiple depths	Infiltration capacity and runoff potential
Impact Observatory LULC (Karra et al., 2021)	10 m annual land cover classification	Surface roughness, imperviousness, and drainage constraints
HydroATLAS (Linke et al., 2019)	Basin-scale attributes (area, climate, land cover summaries)	Static context for streamflow modeling
OPERA DSWx-S1 (Laboratory, 2023)	30 m SAR-based surface water extent	Training target; cloud-independent water detection

Table 1. Summary of data sources. All datasets are globally available and freely accessible, enabling deployment without proprietary data dependencies.

2.2.1 Watershed Delineation and HydroATLAS Attributes

For each study site, we delineate a local "subwatershed" and its corresponding upstream "river basin" using HydroBASINS level-12 polygons from the BasinATLAS v10 geodatabase (Linke et al., 2019). A river basin is defined as the identified subwatershed and all upstream subwatersheds, which are then collected and dissolved into a single "river basin" polygon.

HydroATLAS attributes (including watershed area, built environment metrics, climatology and discharge metrics, lake coverage indices, and land-cover summaries) are aggregated to the river basin polygon using deterministic rules: sums for areas, modes for categorical classes, maxima for hydrologic extremes, and area-weighted means for continuous variables. In the present study, these aggregated HydroATLAS attributes serve primarily as static inputs to the LSTM-FiLM streamflow model (Section 2.3), providing basin-scale context for discharge estimation.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

2.2.2 Dynamic Meteorological Forcing from ERA5

Dynamic meteorological forcings are extracted from daily ERA5 reanalysis data covering all dates with available OPERA SAR-based flood observations (approximately 14 months from September 2024 to November 2025). For each delineated subwatershed and its corresponding upstream river basin, we compute spatially averaged time series for a fixed set of variables: 2-m temperature (mean, minimum, maximum), 2-m dew point temperature, surface solar radiation downwards, and total precipitation.

To perform spatial averaging, we reproject watershed polygons to the native ERA5 coordinate system (0.25° latitude-longitude grid), compute a bounding box around each polygon, and apply a vectorized point-in-polygon mask to the ERA5 grid. For each day, we compute the unweighted mean over all grid cells whose centroids fall inside the polygon. If no cells are found (possible for very small subwatersheds, though rare at level-12 scale), we fall back to nearest-neighbor extraction at the polygon centroid. The result is a pair of daily time series per location—one representing local subwatershed forcing and one representing the aggregated forcing over the entire upstream basin.

These ERA5 timeseries are then cleaned and regularized to provide consistent daily inputs to both the streamflow and inundation models. Missing days within the record (fewer than 0.1% of all daily values) are filled through linear interpolation to form a continuous daily sequence. Given the low frequency of missing data and the predominantly short duration of individual gaps, linear interpolation introduces negligible bias in the resulting daily forcing records. For each watershed type (subwatershed and river basin), we obtain a single, gap-free, multi-variable daily forcing record.

In addition to being used directly by the ML inundation models, these daily forcings also serve as inputs to the LSTM-FiLM streamflow model, which produces daily runoff depth predictions. These are scaled by contributing area to obtain volumetric discharge ($km^3\ day^{-1}$) at both subwatershed and basin scales. From these daily series of streamflow and meteorological variables, we additionally derive simple multi-day aggregations (e.g., 3-day, 7-day, and 14-day trailing sums of precipitation and means of temperature) to provide compact representations of antecedent conditions for models that do not explicitly capture temporal dependence.

2.2.3 Static Geospatial Covariates

Static geospatial covariates are prepared for each subwatershed by clipping and harmonizing DEM, LULC, soil, and HAND products onto a common 30 m resolution grid. This resolution was chosen to match the native resolution of OPERA SAR-based flood masks and to balance computational tractability with the ability to resolve local topographic and land-cover features relevant to inundation processes.

2.2.4 Terrain Metrics from NASADEM

For each subwatershed, we derive a suite of hydrologically relevant terrain metrics designed to capture local topographic position and drainage context at multiple spatial scales (Gnann et al., 2025). Window sizes for all derived terrain and land-cover features were selected to be minimally overlapping across data products, maximizing the diversity of spatial scales represented

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

while limiting features to hydrologically relevant physical scales (from local pixel resolution at 30 m to approximately ∼1 km). This multi-scale design is motivated by the observation that inundation is governed by processes operating across a range of spatial scales: local drainage and microtopography at the pixel level, and landscape connectivity and flow accumulation at intermediate scales (∼1 km). Restricting features to a single spatial scale would force the model to rely on one perspective, potentially missing critical context. The design allows models maximum flexibility to learn which spatial scales are most informative for inundation prediction in different hydrogeomorphic settings, while avoiding excessive redundancy that could arise from densely overlapping window sizes.

From NASADEM, we compute:

- **Elevation anomalies relative to local minimum**: Computed within moving windows of sizes 3×3, 7×7, 15×15, and 31×31 cells (approximately 90 m to 930 m), representing "height above local low point." These features capture an individual pixel's elevation relative to nearby potential flow paths or ponding areas.
- **Inverse anomalies relative to local maximum**: Computed within 5×5, 11×11, 21×21, and 33×33 windows, capturing local depressions or "depth below local high point." These features identify topographic lows where water may accumulate.
- **Anomalies relative to local median**: Computed within 7×7, 13×13, 25×25, and 35×35 windows, identifying elevation anomalies relative to their average surroundings.
- **Upslope area ratio**: The fraction of cells in a moving window that have higher elevation than the focal cell, computed at multiple window sizes. This purely geometric metric serves as a proxy for local drainage convergence versus divergence, without requiring explicit flow direction computation.

These derived DEM features are first computed on a full mosaic of intersecting DEM tiles, then clipped to the subwatershed boundary and resampled via bilinear interpolation to the final 30 m reference grid covering the subwatershed.

2.2.5 Height Above Nearest Drainage

HAND is obtained from the GLO-30 HAND dataset (Nobre et al., 2011), a global 30 m product derived from COPDEM and available through the AWS Registry of Open Data[1]. Tiles are mosaicked where necessary, clipped to the watershed boundary, and resampled via bilinear interpolation onto the 30 m reference grid. HAND provides an estimate of the vertical distance from each pixel to its nearest drainage channel, serving as a first-order proxy for inundation susceptibility (Nobre et al., 2011). By incorporating HAND derived from COPDEM alongside elevation metrics derived from NASADEM, we leverage complementary information from both DEM products, allowing models to implicitly account for regional biases in either dataset.

[1] https://registry.opendata.aws/glo-30-hand/

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

2.2.6 Soil Properties from SoilGrids

Soil texture properties (clay, sand, and silt mass fractions) are retrieved from SoilGrids 2.0 (Hengl et al., 2017) at two non-overlapping depth intervals per texture class, selected to span the full soil profile while minimizing feature redundancy. Specifically, we extracted percent clay (0–5 cm, 30–60 cm), percent sand (5–15 cm, 60–100 cm), and percent silt (15–30 cm, 100–200 cm), clipped to the watershed polygon, and resampled via bilinear interpolation to the 30 m DEM grid. SoilGrids NoData values (typically indicating water bodies, rock outcrops, or ice) are explicitly handled and preserved. Post-resampling, we validate that resulting values fall within physically plausible ranges (0–100% for texture fractions) and flag any anomalies for quality control.

2.2.7 Land Use and Land Cover

Land use and land cover are represented using the Impact Observatory 10 m annual global LULC product (Karra et al., 2021), a Sentinel-2 derived classification covering nine land-cover classes. For each subwatershed, we query and mosaic all intersecting LULC tiles, clip them to the watershed polygon, and derive additional neighborhood-based features to characterize the spatial context around each pixel. These neighborhood features are computed at the native 10 m resolution, then resampled to the 30 m DEM reference grid via nearest-neighbor interpolation to preserve categorical (modal class) and fractional (built-up ratio) characteristics. The base LULC classification is resampled in the same manner:

- **Modal LULC class**: Computed within sliding windows of increasing size (30-m to 1-km), representing the dominant land-cover context.
- **Built-up area fraction**: Computed as the ratio of pixels classified as built-up or urban within 90 m, 270 m, and 810 m extents. This serves as a proxy for local urbanization intensity and potential drainage constraints due to impervious surfaces and stormwater infrastructure (Leopold, 1968).

2.2.8 Static Feature Stack Assembly

Finally, all DEM-based metrics (base elevation plus 14 derived terrain features), HAND, soil layers (three texture classes at two depths each), LULC base classification, modal LULC at four scales, and built-up fraction at three scales are stacked into a single multi-band GeoTIFF per subwatershed at 30 m resolution. The resulting static feature stack (Figure 2) comprises 30 bands per subwatershed, and each band is assigned consistent georeferencing, NoData handling, and metadata tags.

2.2.9 OPERA DSWx-S1 Ground Truth

To provide a consistent and analysis-ready ground truth for surface water extent, we utilize the OPERA Dynamic Surface Water Extent from Sentinel-1 (DSWx-S1) product (Laboratory, 2023). DSWx-S1 delivers near-global, 30 m resolution surface water classifications derived from SAR data acquired by Sentinel-1A/B satellites, mapped on the Military Grid Reference System (MGRS) tiling scheme. The product is specifically designed to detect open inland water bodies larger than 3 hectares and at

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

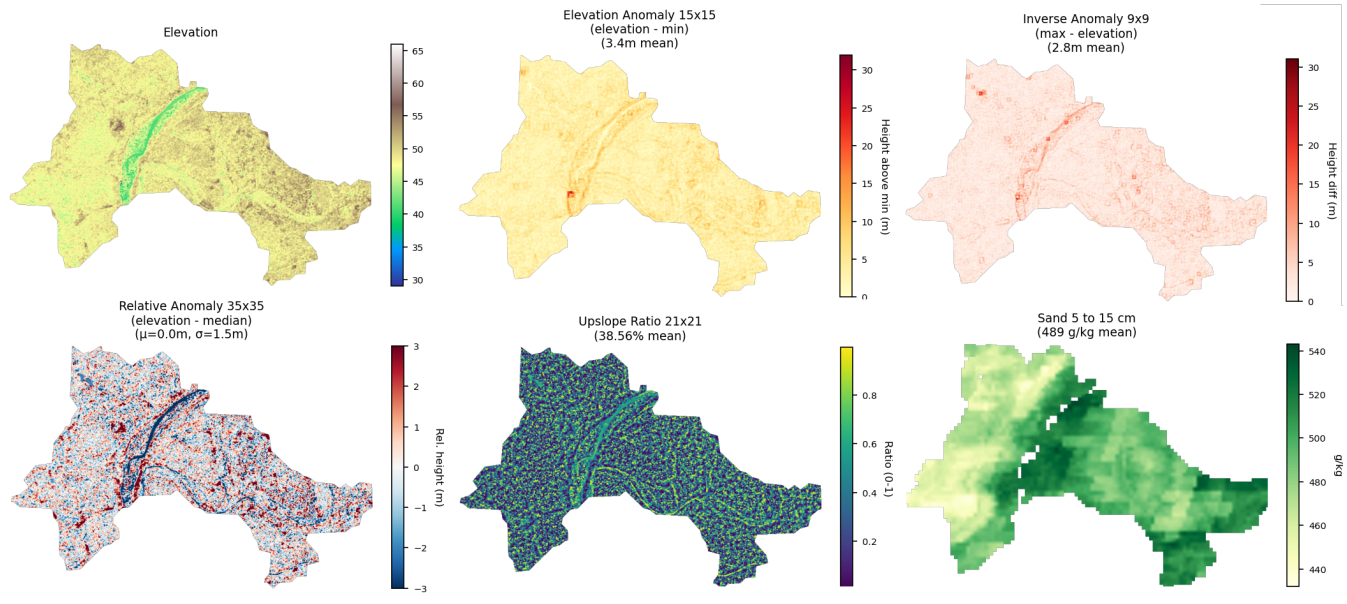

Figure 2. Example (target subwatershed of location 7 (Pakistan)) of static geospatial datasets and derived features used in model training. Six example features are illustrated here clockwise from top left: 1) elevation (relative to training subwatershed minimum); 2) Elevation Anomaly within a 15x15 grid (local elevation over nearby minima); 3) Inverse Anomaly over a 9x9 grid (local elevation below nearby maxima); 4) Sand 5 to 15 cm (percent sand at 5 to 15 cm depth profile); 5) Upslope Ratio within a 21x21 grid (percent of nearby land parcels that are upslope of a local point); and 6) Relative Anomaly within a 35x35 grid (local elevation relative to nearby averages). An example of all features is given in the appendix.

least 200 m in width, with a temporal revisit frequency of 6–12 days, enabling robust monitoring of large-scale flood events and riverine inundation independent of cloud cover or daylight conditions. The DSWx-S1 workflow applies thresholding and contextual algorithms to radiometrically terrain-corrected Sentinel-1 backscatter, with documented user accuracy of approximately 86% and producer accuracy of 94% when validated against higher-resolution remote sensing-based reference datasets, suggesting an F1 of roughly 0.89. However, the product exhibits reduced sensitivity to small or narrow water bodies (below 3 hectares or 200 m width), and may be challenged by mixed pixels, emergent vegetation, or complex urban environments. These limitations are particularly relevant for our study regions, which intentionally feature small, fragmented water bodies. Despite these constraints, DSWx-S1 provides a standardized, analysis-ready benchmark for hydrological model training (Laboratory, 2023).

2.3 Streamflow Modeling

Due to the lack of globally available watershed-scale streamflow discharge data, we first develop a simple scalable streamflow model that depends exclusively on global open-source climate forcings and watershed attributes. The streamflow model is an LSTM (Kratzert et al., 2019) with feature-wise linear modulation (FiLM) (Perez et al., 2018) of hidden states by static

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

basin attributes. Dynamic inputs (daily meteorological forcings and derived precipitation/snow variables) are fed through a stacked LSTM with two layers, hidden size of 256, dropout of 0.5, learning rate of 0.0005, and 40 epochs with early stopping. Static catchment descriptors (primarily HydroATLAS attributes and basin area) are passed through two separate multilayer perceptrons: one initializes the LSTM's hidden state, and the other is a FiLM mechanism that produces a vector of "modulation logits." These logits are transformed with a sigmoid and applied multiplicatively to the LSTM hidden states at every time step, yielding a basin-specific FiLM conditioning on static attributes. The final discharge prediction is obtained by concatenating the last time-step hidden state with the static feature vector and passing this combined representation through a small fully connected readout network.

Training data were prepared by aggregating daily time series across gauged watersheds ($n > 2000$) from the Caravan database (Kratzert et al., 2023) into a single "pooling" dataset and then constructing fixed-length sequences for supervised learning. For each watershed, we loaded daily streamflow and ERA5-derived forcings, derived additional hydrometeorological features (e.g., 5-day smoothed and 90-day lagged precipitation, simple temperature-based rain–snow partitioning), and filled gaps to obtain continuous daily records. Static HydroATLAS attributes were attached to every daily row within a watershed and later standardized (z-scored) across the training set. Dynamic predictors were normalized using a combination of z-score and min–max scaling, while streamflow targets were first transformed with log1p and then standardized using the training-set mean and standard deviation. This design allowed a single LSTM-FiLM model to learn from many heterogeneous basins while conditioning explicitly on their static attributes.

Formally, the FiLM conditioning mechanism operates as follows. Given dynamic inputs x_t at time step t and static catchment descriptors s, the LSTM hidden state h_t is computed as:

$$h_t = \mathrm{LSTM}(x_t, h_{t-1})$$

Two separate MLPs process the static attributes to produce modulation parameters:

$$\gamma = \sigma(\mathrm{MLP}_\gamma(s)), \quad \beta = \mathrm{MLP}_\beta(s)$$

where σ denotes the sigmoid function. The modulated hidden state is then:

$$\tilde{h}_t = \gamma \odot h_t + \beta$$

where $\odot$ denotes element-wise multiplication. The final discharge prediction is obtained by concatenating the last time-step modulated hidden state with the static feature vector and passing this combined representation through a fully connected readout network. Predicted runoff depth $\hat{d}_t$ is converted to volumetric discharge via:

$$Q_t = \hat{d}_t \times A$$

where A is the contributing watershed area.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Model training and tuning were managed using PyTorch Lightning (Falcon, 2019). We split basins into training, validation, and test sets using a PCA-stratified procedure on HydroATLAS attributes to ensure that each subset spanned the range of hydroclimatic and physiographic conditions, and we then constructed sequence datasets for each split. We chose a composite loss function that prioritizes minimizing error during high flows (i.e., floods) rather than a traditional balanced hydrologic metric (e.g., NSE or KGE):

$$\mathcal{L} = \text{MAE}(y, \hat{y}) + \text{RMSE}(y, \hat{y}) + \left(\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|^3 \right)^{1/3}$$

This composite loss balances complementary optimization objectives: MAE weights all errors equally, anchoring predictions to typical flow conditions; RMSE penalizes large errors quadratically, increasing sensitivity to extreme events; and the cube-root of mean absolute cubed error (RMCE) amplifies this emphasis further by penalizing large errors cubically, placing the greatest weight on the highest-magnitude deviations. Together, these terms create a progressive weighting scheme that ensures the model performs well across typical conditions while increasingly prioritizing accuracy during rare extreme flows, which is the regime most critical for flood forecasting applications. We used the Adam optimizer (Kingma and Ba, 2014) with a fixed learning rate (0.0005) with gradient clipping and a ReduceLROnPlateau scheduler that halved the learning rate when validation loss stagnated. After 40 epochs with a 60/20/20 train/validate/test split, the best checkpoint generated test results of median KGE = 0.62, NSE = 0.63, MAE = 0.49, RMSE = 1.17, RMCE = 1.82, and r^2 = 0.74. We also calculated F1 scores for n-year events (with success being defined as a prediction of an n-year event within 2 days of an observed n-year event, with n-year events estimated according to the Log-Pearson Type III distribution). As a preliminary validation, our LSTM-FiLM model achieved F1 scores of 0.65 (2-year flood), 0.56 (5-year flood), and 0.50 (10-year flood), compared to 0.35, 0.42, and 0.40 for Google Flood Hub. Together, these performance metrics indicate substantial per-watershed bias in absolute values but strong correlation with extreme event timing in our LSTM-FiLM model, which fits our design criteria.

However, we stress that the BAGIM framework is agnostic to the source of streamflow predictions, and therefore any streamflow model (global or local, physics-based or ML) may be used in replacement of the LSTM-FiLM model described here, without requiring the streamflow model to be explicitly specified as part of the framework.

2.4 Flood Inundation Model Architectures

2.4.1 Model Selection Rationale

We evaluate six model architectures spanning a spectrum of complexity to test whether increased architectural sophistication improves inundation prediction beyond what can be achieved through hydrologically meaningful feature engineering alone. The six architectures are:

- **Generalized Linear Model (GLM)**: Regularized logistic regression acting on standardized pixel-wise feature vectors (static and temporally varying predictors concatenated) (Nelder and Wedderburn, 1972). Provides a transparent, linear-

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

in-the-features baseline. Tuned hyperparameters: regularization strength (C), penalty type (L1, L2, or elastic net), solver (liblinear or saga), and maximum iterations.

- **Random Forest**: An ensemble of decision trees trained on standardized pixel-wise feature vectors (Breiman, 2001). Tuned hyperparameters: number of trees (100–500), maximum depth (10–30 or unlimited), minimum samples per split (2–10), minimum samples per leaf (1–4), and feature sampling fraction at each split.

- **XGBoost**: Gradient-boosted trees optimized for large-scale, imbalanced tabular data using histogram-based tree construction (Chen and Guestrin, 2016). Tuned hyperparameters: learning rate (0.01–0.3), maximum depth (3–10), number of estimators (100–500), subsampling fractions (0.6–1.0), minimum child weight (1–7), and L1/L2 regularization parameters. Early stopping (patience 20 rounds) is applied based on validation log-loss.

- **TabNet**: A deep-learning alternative for tabular data with built-in attentive feature selection via learnable attention masks (sparsemax or entmax) across multiple decision steps (Arik and Pfister, 2021). The architecture selects relevant features at each layer, providing interpretability complementary to tree-based methods. Tuned hyperparameters: decision and attention dimensions (n_d, n_a), number of steps (3–6), gamma (relaxation parameter), and sparsity regularization (λ_{sparse}). Training uses Adam optimization with early stopping based on balanced accuracy.

- **Multi-Layer Perceptron with FiLM (MLP-FiLM)**: A pixel-wise feed-forward neural network (Rumelhart et al., 1986) that explicitly separates static and temporal predictors, using feature-wise linear modulation (FiLM) to condition hidden representations on temporally varying, basin-consistent forcings (sub-watershed and basin-aggregated streamflow and weather). The FiLM mechanism applies learned scale and shift parameters derived from temporal inputs to modulate static feature representations at designated hidden layers. A learned interaction module projects static features into a lower-dimensional space and gates these projections with temporal inputs, yielding "interaction features" that capture how static landscape properties modulate current hydrometeorological conditions. Training uses AdamW (Llugsi et al., 2021) with focal loss and mixed-precision training on GPU.

- **U-Net with FiLM (U-Net-FiLM)**: Extends FiLM conditioning into the spatial domain, operating on 2D tiles of static raster features while conditioning on temporally varying basin-scale predictors. Static inputs consist of multi-band image stacks (DEM-derived indices, HAND, soils, land cover, neighborhood statistics) cropped to fixed-size tiles (128–256 pixels). These are processed by a standard U-Net encoder-decoder (Ronneberger et al., 2015): an encoder path progressively reduces spatial resolution while increasing channel depth; a bottleneck captures global context; and a decoder path reconstructs high-resolution feature maps via transposed convolutions and skip connections. Each convolutional block includes a FiLM module that generates channel-wise scale and shift parameters from per-tile temporal forcings, allowing temporal conditions to modulate spatial feature extraction at multiple scales. Training uses a composite loss combining masked binary cross-entropy (ignoring SAR NoData pixels) and masked Dice loss (Milletari et al., 2016) to balance precision and recall.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

These six architectures span three broad categories: a transparent linear baseline (GLM); tree-based ensembles (Random Forest, XGBoost) that capture non-linear interactions without spatial structure; and deep learning architectures of increasing complexity (TabNet, MLP-FiLM, U-Net-FiLM) that incorporate attention-based feature selection and FiLM-based temporal or spatial conditioning. All six are trained on identical engineered feature sets under the same watershed-based cross-validation, adaptive sampling strategy, threshold selection procedure, and hyperparameter tuning protocol, enabling direct comparison across the complexity spectrum while controlling for feature engineering, training protocol, and evaluation methodology. If simpler models perform comparably to more complex architectures, this would support the hypothesis that feature engineering, rather than model complexity, is the primary driver of predictive skill in this domain.

2.4.2 Common Training Framework

Across all six inundation models, we adopt a consistent training and evaluation protocol to ensure comparability and isolate the effect of architectural complexity. For each basin-specific dataset, we compute feature normalization statistics (means and standard deviations) using only training watersheds, then standardize all numeric predictors accordingly. Watersheds are partitioned spatially (Kratzert et al., 2021) via leave-one-watershed-out cross-validation with a fixed held-out test watershed. To address severe class imbalance, we apply adaptive per-watershed sub-sampling during training: all positive ("water") pixels are retained, while negative pixels are subsampled to achieve a target positive fraction determined by natural flood prevalence using square-root compression (Yan et al., 2015; Chawla et al., 2002; He and Garcia, 2009). Square-root compression moderates the degree of class rebalancing: rather than equalizing class frequencies (which can overweight rare events) or preserving natural frequencies (which can underweight them), this approach applies a concave transformation that partially corrects imbalance while retaining information about relative flood prevalence across watersheds. This per-watershed calibration preserves local hydrological signals rather than imposing a uniform class distribution that could obscure meaningful regional variation. Models are trained on aggregated samples from training watersheds and evaluated on validation watersheds. For all models, decision thresholds are determined by maximizing F1 score on the validation precision-recall curve, and final thresholds are averaged across cross-validation folds. Hyperparameters for each model family are tuned using Optuna (Akiba et al., 2019; Snoek et al., 2012) with 20 trials, and the resulting best configuration is applied in all basin-level experiments.

Figure 3 illustrates the general workflow for our approach.

2.4.3 Evaluation Metrics

Model performance is evaluated at the pixel level by comparing each predicted binary inundation mask against the coincident OPERA DSWx-S1 reference snapshot. For every (basin, date) pair in the held-out set we tabulate true and false positives and negatives and compute seven complementary skill metrics: overall pixel Accuracy; Precision and Recall; their harmonic mean (F1); the Matthews Correlation Coefficient (MCC), a correlation-based score that is well-behaved under severe class imbalance (Chicco and Jurman, 2020; Powers, 2011); and the areas under the Receiver Operating Characteristic and Precision-Recall curves (ROC-AUC and PR-AUC), which summarize the underlying probabilistic classifier in a threshold-independent way. Because water pixels typically constitute less than 5% of each scene, Accuracy is almost always very high (> 0.95) and

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

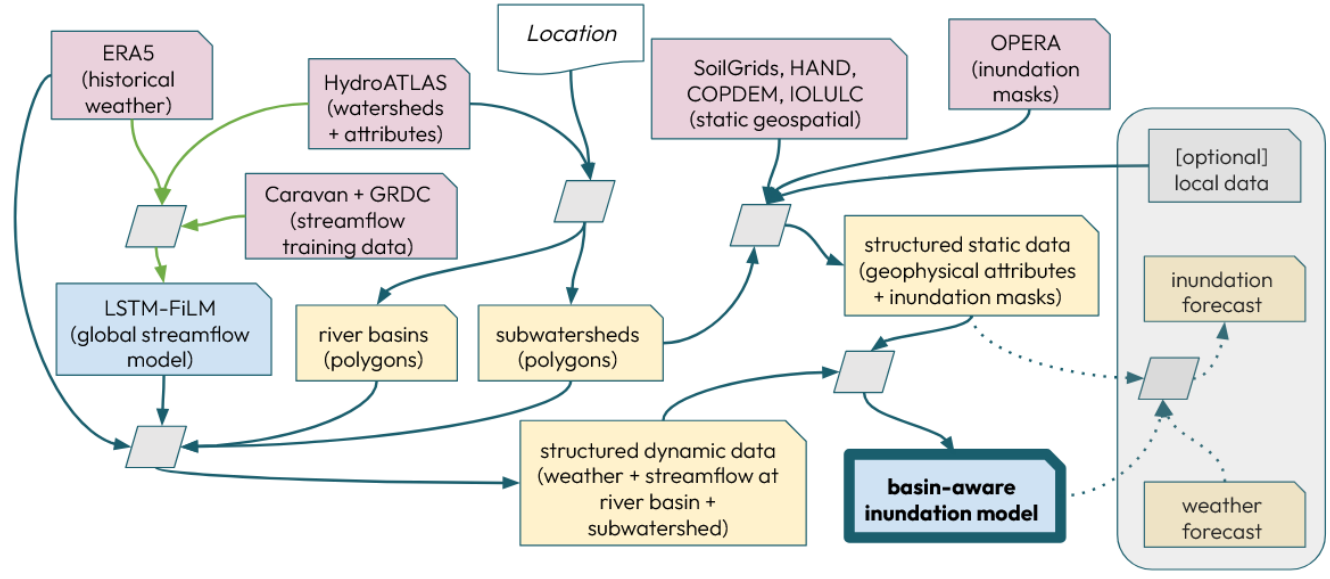

Figure 3. BAGIM workflow overview. **Inputs** (red): Raw data inputs are highlighted in red boxes. The specific data inputs used in this manuscript are noted in each box, but we reiterate that within the BAGIM framework these specified sources may be replaced or supplemented with alternatives. **Processing** (gray): Data processing and parameter tuning are indicated by gray rhombuses. **Models** (blue): Model final states (architecture and hyperparameter values) are highlighted in blue. **Outputs** (yellow): Processed data outputs (some of which are themselves inputs for subsequent steps of the framework) are highlighted in yellow. **Workflows** (arrows): Green arrows indicate steps that are taken only once (to generate a global model), blue arrows indicate primary steps to develop a basin-aware inundation model for a target watershed, and dotted lines indicate additional steps needed to operationalize an inundation forecast. Operationalization of forecasts and inclusion of local data are beyond the scope of this paper, as indicated by the gray box obscuring these steps.

of limited diagnostic value; we therefore rely on F1 and MCC as the primary indicators of model skill, and F1 is also the score against which decision thresholds are tuned during training. All metrics are computed separately for each (basin, date) pair and aggregated in the results figures; pairwise between-model comparisons use the Mann-Whitney U rank-sum test at $\alpha = 0.05$.

2.4.4 Experimental Design for Hypothesis Testing

375 The experiments are designed to test our three hypotheses.

H1 – Feature engineering over complexity. H1 is tested in two stages. First, we compare performance across the six architectures of varying complexity described above (Figure 5). Second, we run four XGBoost feature- and target-manipulation experiments (Figure 7):

- **XGBoost Basic Features**: Only includes coincident river basin streamflow, HAND, and elevation. Intended to assess
380 model performance when only including the features commonly used in popular operational systems like Google Flood Hub (Hub, 2022) and GloFAS (Alfieri et al., 2013).

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

- **XGBoost No Streamflow**: Explicitly excludes river basin and subwatershed streamflow as features and only includes weather and static geospatial features. Intended to test whether or not streamflow is essential to estimate inundation extent or whether an overlying model is capable of inferring streamflow impacts when supplied well-designed weather features.
- **XGBoost Multi Objective**: All "raw" features (local LULC, elevation, and coincident streamflow) are initially removed from the feature set, then each is individually predicted by the aggregated features with independent XGBoost models, then all derived and estimated features are used in the feature set for final model training. Intended to assess whether smoothing specificity in the underlying features impacts model performance.
- **XGBoost 25% Noise**: The training target is manipulated by randomly selecting 25% of all positives (pixels inundated by water) to be flipped to negative, and the same number of negatives randomly selected to be flipped to positive. Intended to assess whether increased uncertainty in the "ground truth" dataset degrades a model's capacity to make realistic hydrological inferences from the remaining signal.

H2 – Basin-scale training mitigates regional bias. We train one baseline XGBoost model per basin and evaluate all 56 cross-basin transfer scenarios (each of the eight basin-trained models deployed against the other seven basins), quantifying the degradation in held-out-basin skill metrics relative to the corresponding in-basin model (Figure 8).

H3 – Generalization to out-of-sample events. For each of the eight study locations, we identify an extreme flood event from the STURM-Flood dataset (Notarangelo et al., 2025) – a high-resolution benchmark dataset specifically curated for rare, high-magnitude floods – that overlaps spatially but not temporally with our OPERA training window. We then evaluate the XGBoost model (trained exclusively on non-extreme OPERA observations) against both the STURM reference and the coincident Copernicus Sentinel-2 classification (Figures 9 and 10).

Ensemble aggregation. For each experimental family above, we additionally report a pixel-level majority-vote ensemble (ties resolved to positive). Four ensembles are constructed (Figure 11):

- **Architecture Ensemble**: An ensemble of all 6 architectures.
- **Feature Ensemble**: An ensemble of the four XGBoost feature- and target-manipulation experiments.
- **Transfer Ensemble**: An ensemble of the eight per-basin baseline XGBoost transfer models.
- **Kitchen Sink**: An ensemble of all experiments, with redundant XGBoost Baseline examples being eliminated.

3 Results and Discussion

3.1 General Model Performance (H1)

BAGIM produces physically plausible inundation predictions that are consistent with observable hydrological features across diverse hydroclimatic regimes. Figure 4 illustrates this for a representative date (2024-11-20) in Greece (Location 6, a pluvial

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

flash-flood event), where the XGBoost model achieves higher accuracy when validated against Copernicus's S2 inundation masks (F1 = 0.51) than when tested against the OPERA DSWx-S1 reference (F1 = 0.30), despite having been trained exclusively on the latter. Critically, most "false" positives generated by the model are concentrated along river channels with high NDWI values, suggesting that these discrepancies reflect genuine hydrological features rather than model error. This example also highlights the substantial variability among global flood map products (Figure 4): both the XGBoost model and the Copernicus Sentinel-2 (S2) classification product exhibit discrepancies when compared to the OPERA DSWx-S1 reference (F1 = 0.17), particularly for small or fragmented water bodies, which underscores the inherent uncertainty ceiling imposed by remote sensing-derived ground truth.

This analysis highlights a fundamental principle: while performance metrics such as MCC, F1, ROC-AUC, and PR-AUC (Powers, 2011) are essential, model outputs must also be assessed for their physical and hydrological plausibility. No single ground truth or metric can be taken at face value given the substantial uncertainty in remote sensing-derived flood masks (Laboratory, 2023; Jafarzadegan and et al., 2023). Nonetheless, quantitative performance metrics remain our primary tool for assessing model quality, and we proceed with the intuition that aggregating results across large scales (test: 1.6×10^9; train: 1.6×10^{10}) and diverse environments (eight countries) allows for a robust assessment of relative model performance.

Figure 5 presents a comparative analysis of six model architectures (GLM, Random Forest, XGBoost, MLP-FiLM, U-Net-FiLM, TabNet) aggregated across all eight test watersheds. Average model performance is approximately 0.5 for MCC and F1, with Accuracy exceeding 0.95, indicative of strong class imbalance. All model architectures outperform the linear GLM baseline in most metrics, indicating the necessity of capturing non-linear hydrologic responses (Kirchner, 2006; Sivakumar and Singh, 2012). XGBoost and Random Forest consistently achieve the highest performance, while more complex architectures such as U-Net-FiLM and TabNet tend to underperform, potentially due to overfitting given the available data. We note that the approximately 14-month observation window used here may disproportionately limit architectures with greater parameter counts; with substantially longer training records, deep learning architectures may narrow this performance gap, as such models typically benefit more from increased data volume than tree-based methods. Nonetheless, these results suggest that increased model complexity does not guarantee improved accuracy in this context, and that tree-based ensemble methods offer a robust balance of interpretability and predictive skill (Lundberg et al., 2020). Critically for operational deployment, this finding indicates that BAGIM can achieve strong performance without requiring specialized GPU infrastructure or global-scale training, enhancing accessibility for organizations with limited computational resources while maintaining regional sensitivity. Due to the high performance and computational efficiency of XGBoost, we preferentially leverage it for subsequent experiments.

3.2 Feature Engineering and Model Complexity (H1)

Feature importance analysis (Table 2) reveals that derived and smoothed features (i.e., Land Use Land Cover (LULC) in 3×3 and 9×9 grids) are the most influential predictors across model architectures. This suggests that aggregating or smoothing raw data can help mitigate uncertainty and noise, leading to more robust model performance. Cumulative hydrological variables, such as cumulative streamflow over recent weeks, are also ranked higher than coincident river basin and subwatershed streamflow. While Height Above Nearest Drainage (HAND) is ranked highly at fourth, this ranking is lower than the authors'

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

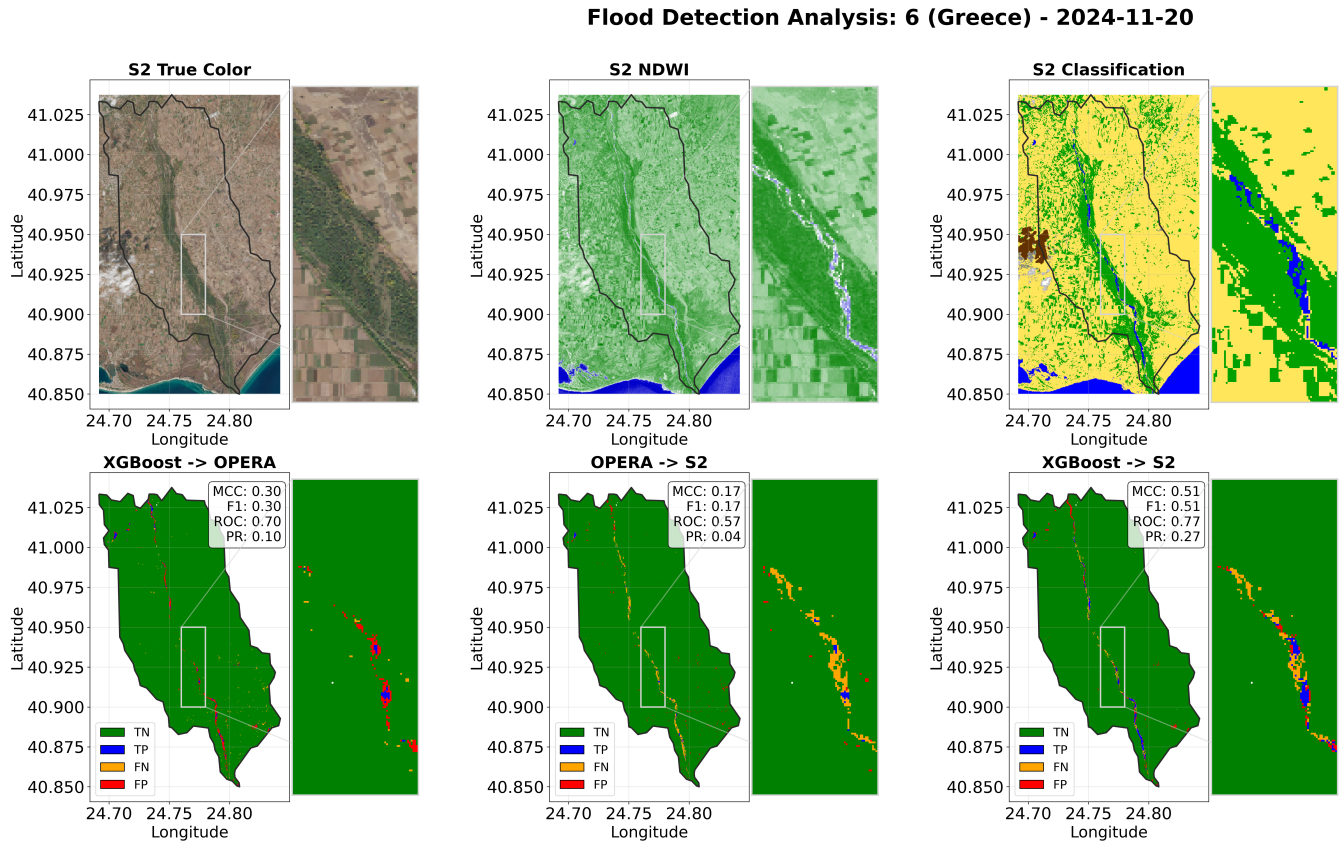

Figure 4. Clockwise from top left: S2 True Color is a true color image of the target subwatershed on 2024-11-20; the inset presents a ”zoomed in” visualization of a small intermittent stream channel running north to south and the abutting riparian zone. S2 NDWI is the normalized difference water index as computed on the image by the Copernicus online portal; the inset of the small river channel is indicated by a narrow width of high NDWI anomalies, indicating low volume intermittent streamflow. S2 Classification, also derived directly from the Copernicus online portal, indicates water in blue; note the identification of isolated patches of surface water. XGBoost -> S2 presents a ”confusion matrix” style map indicating the ”accuracy” of our XGBoost model when predicting Copernicus’s online portal’s water mask; note the false positive (FP) values indicated along the river channel in red. OPERA -> S2 presents OPERA when predicting Copernicus’s online portal’s generic classification map; note the false negative (FN) values indicated along the river channel in yellow. XGBoost -> OPERA presents our XGBoost model when predicting Copernicus’s online portal’s generic classification map; note that false negative and false positive values tend to congregate along the intermittent channel where the underlying ”ground-truth” classifications are most subject to uncertainty.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

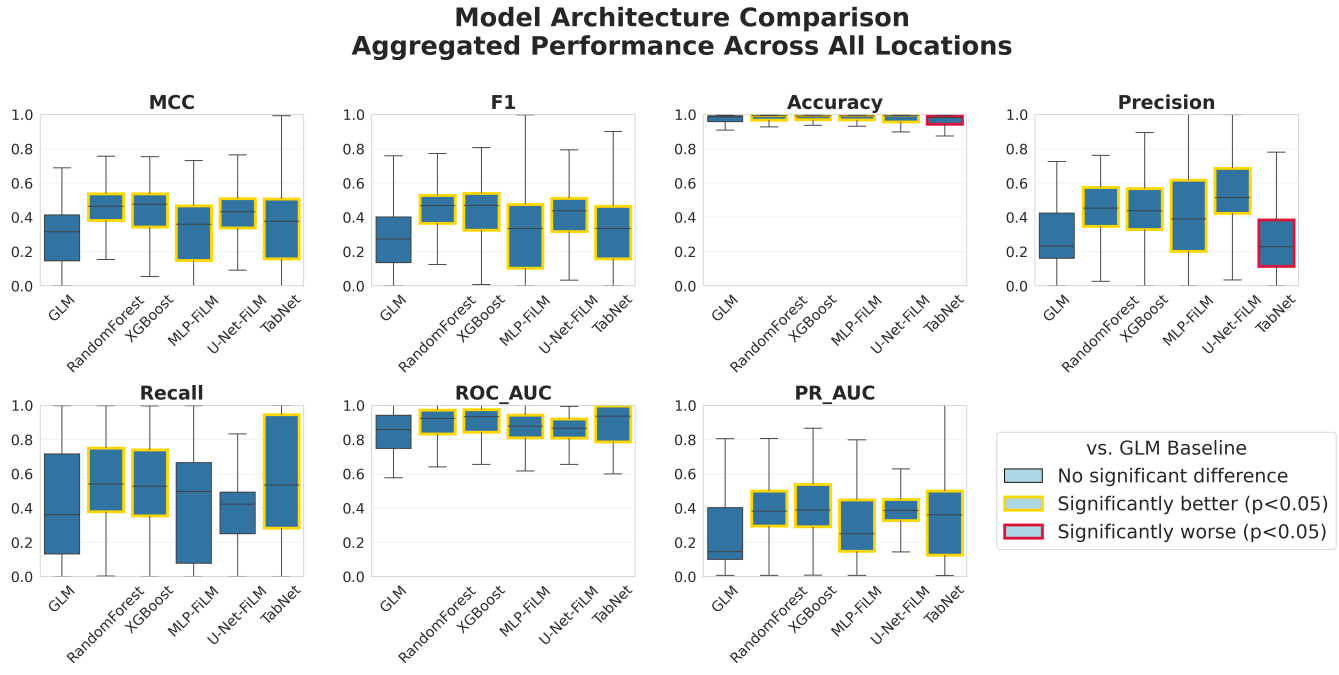

Figure 5. Model architecture comparison across seven evaluation metrics. Tree-based ensembles (XGBoost, Random Forest) consistently outperform more complex deep learning architectures (U-Net-FiLM, TabNet), supporting H1. Boxplots represent aggregate scores across all 8 target locations and all dates. Metrics shown (clockwise from top left): MCC (Matthews Correlation Coefficient), F1, Accuracy, Precision, PR-AUC (Precision-Recall Area Under Curve), ROC-AUC (Receiver Operating Characteristic Area Under Curve), and Recall. Models highlighted in gold or red indicate statistically significant differences ($p < 0.05$) from the GLM baseline, computed via Mann-Whitney U test. Note that accuracy exceeds 0.95 for all models due to strong class imbalance.

expectations given that it is a feature that is used directly in the derivation of the OPERA dataset (Laboratory, 2023) perhaps due to collinearity with NASADEM derived features. Notably, several features exhibit high maximum importance scores but low mean scores (reflected in high IQR values). This pattern suggests basin-specific relevance: certain topographic features may be highly predictive in specific hydroclimatic contexts while contributing little in others (Figure 6), reinforcing the value of basin-aware calibration that can leverage locally relevant predictors.

Figure 7 presents the results of the four XGBoost feature- and target-manipulation experiments defined in Section 2.4.4. XGBoost Basic Features performs very poorly, indeed substantially worse than all other experiments in this analysis in nearly every category. Worryingly, this implies that the standard baseline features used in flood prediction models are inadequate, at least in the context of predicting inundation during non-extreme events. Conversely, XGBoost No Streamflow performs comparably to the baseline model (which does include river basin and subwatershed streamflow), indicating that well-engineered meteorological inputs provide sufficient context for ML models to infer inundated areal extent, at least during non-extreme events. These results suggest an alternative to the paradigm of many ML-based operational flood model architectures, indi-

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Feature	Mean	Median	Max	IQR
most_common_lulc_neighbor_3x3	0.1594	0.1118	0.4026	0.1992
most_common_lulc_neighbor_9x9	0.0952	0.0700	0.4741	0.0787
IO_LULC	0.0902	0.0522	0.6167	0.1063
HAND	0.0569	0.0456	0.3018	0.0556
elevation	0.0595	0.0425	0.1745	0.0287
elevation_anomaly_31x31	0.0489	0.0117	0.4075	0.0146
SAND_5_TO_15CM	0.0174	0.0102	0.0742	0.0301
most_common_lulc_neighbor_27x27	0.0129	0.0091	0.0669	0.0175
CLAY_30_TO_60CM	0.0207	0.0079	0.1268	0.0275
CLAY_0_TO_5CM	0.0148	0.0078	0.0531	0.0240
SILT_100_TO_200CM	0.0153	0.0077	0.0713	0.0181
SAND_60_TO_100CM	0.0138	0.0071	0.0467	0.0206
elevation_anomaly_7x7	0.0522	0.0071	0.4042	0.0085
elevation_anomaly_15x15	0.0298	0.0066	0.1958	0.0088
elevation_anomaly_inverse_33x33	0.0105	0.0066	0.0266	0.0191
river_basin_streamflow_pred_32	0.0079	0.0059	0.0446	0.0101
SILT_15_TO_30CM	0.0112	0.0057	0.0397	0.0187
river_basin_streamflow_pred_16	0.0072	0.0053	0.0258	0.0107
river_basin_t2m	0.0073	0.0053	0.0327	0.0069
subwatershed_streamflow_pred_32	0.0053	0.0050	0.0162	0.0077

Table 2. Top 20 features ranked by importance across all six model architectures. Mean, median, maximum, and interquartile range (IQR) of feature importance scores computed across six architectures, eight basins, and ten cross-validation folds (N = 480 model instances). For tree-based models (Random Forest, XGBoost), importance reflects mean decrease in Gini impurity. For TabNet, importance is derived from aggregated attention mask weights. For neural models (MLP-FiLM, U-Net-FiLM), importance is computed via gradient-based saliency scores.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

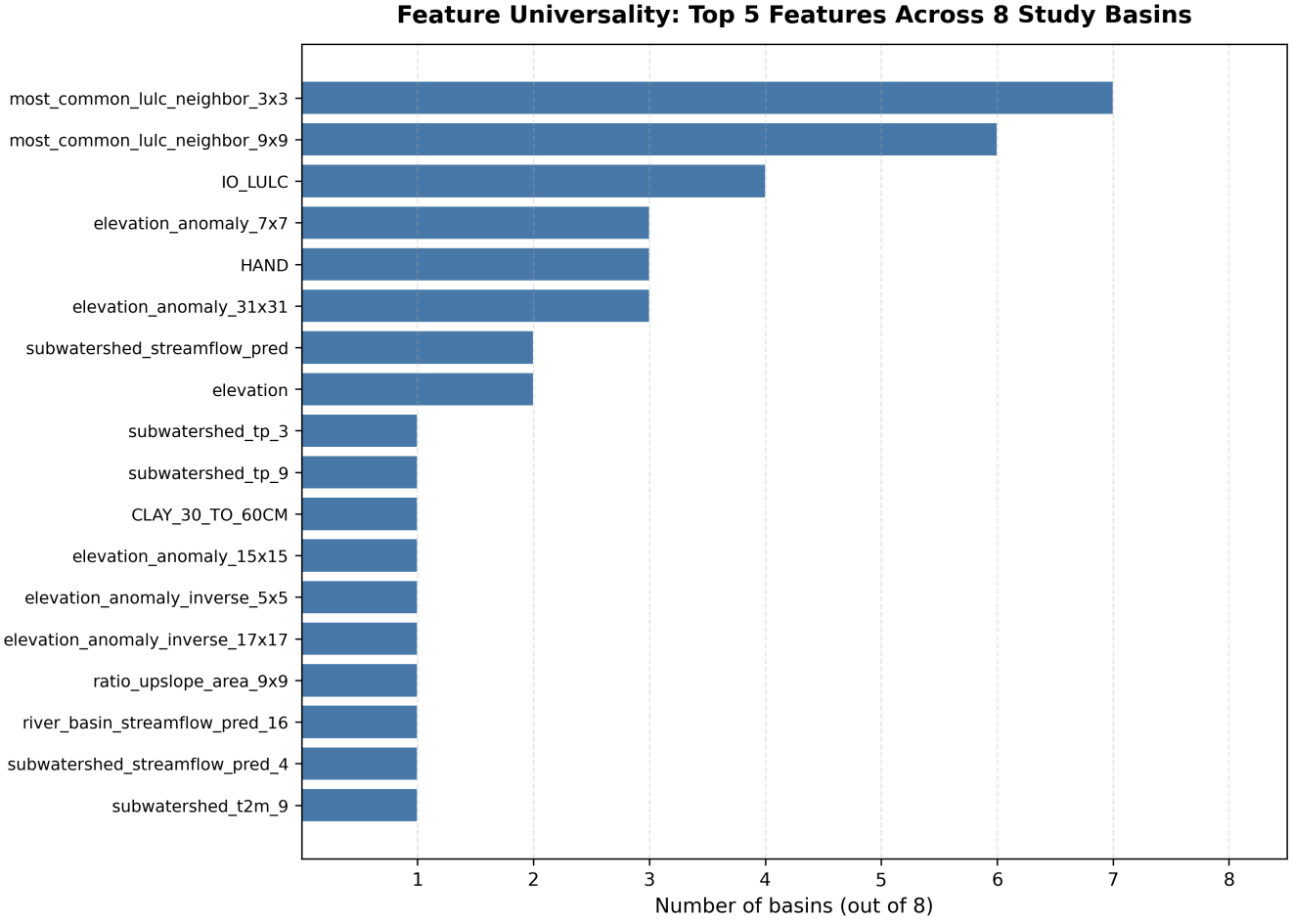

Figure 6. Feature universality across the eight study basins. Each bar shows the number of basins for which a given feature ranks among the top five by mean normalized importance. LULC features dominate: modal LULC in a 3×3 neighborhood appears in seven of eight basins (all except Togo), and modal LULC in a 9×9 neighborhood appears in six (all except Togo and Madagascar), and local LULC (IO LULC) ranks in the top five for Greece, Spain, Uganda, and the United Kingdom. Terrain-related features show more regional specificity: HAND appears in three (Pakistan, Togo, and Ukraine) as does elevation anomaly at 31×31 scale (Pakistan, Uganda, and Ukraine) and elevation anomaly at 7×7 scale (Madagascar, Togo, and Ukraine), while elevation appears in two (Greece and Togo). Streamflow predictions appear for Madagascar and Spain (subwatershed-scale) and Spain alone (river-basin-scale 16-day average). Climate variables enter the top five only for Pakistan (subwatershed precipitation, 3-day), the United Kingdom (subwatershed precipitation, 9-day average), and Madagascar (subwatershed temperature, 9-day average). The remaining basin-specific features are clay content at 30–60 cm (Togo), upslope contributing area ratio at 9×9 scale (Uganda), inverse elevation anomaly at 5×5 scale (Greece), and inverse elevation anomaly at 17×17 scale (the United Kingdom). The remaining 38 features are not shown because they do not rank in the top five for any basin.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

cating that BAGIM models can achieve robust inundation predictions without requiring an upstream streamflow model, thus reducing system complexity and eliminating a potential source of cascading error.

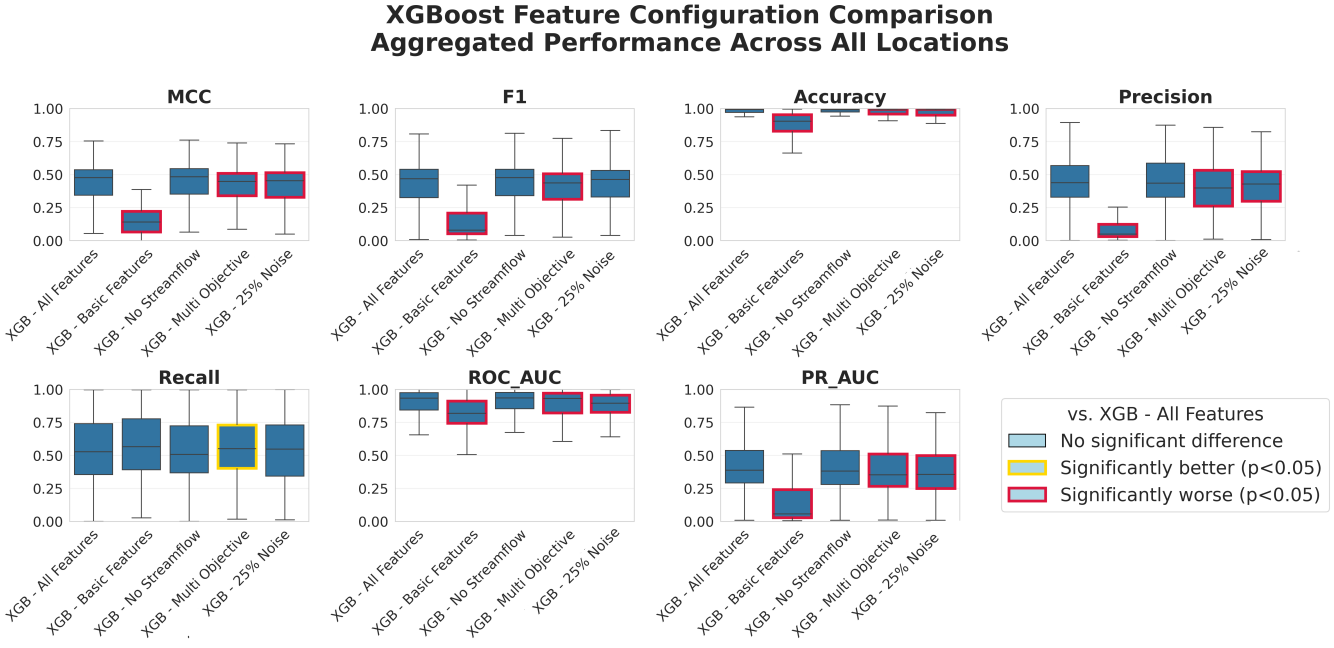

Figure 7. Results of feature and target manipulation experiments evaluated on held-out target watersheds (out-of-sample in space), comparing the baseline XGBoost model (XGB - All Features) to an XGBoost model with only coincident river basin streamflow, elevation, and HAND as features (XGB - Basic Features), an XGBoost model without streamflow as a feature (XGB - No Streamflow), an XGBoost model with primary features estimated from derived features (XGB - Multi Objective), and an XGBoost model where 25% of water labels have been randomly flipped. For each of eight study locations, models are trained and validated on 10 proximal watersheds and tested on the target watershed. Models are highlighted Gold or Red to indicate statistically significant ($p < 0.05$) differences from the XGB - All Features baseline model, as computed by the Mann-Whitney U test.

Promisingly, XGBoost 25% Noise performs nearly as well as the baseline model, with a statistically significant yet numerically modest reduction in skill (median F1 and MCC decrease by 0.01 and 0.02, respectively). This resilience to substantial label corruption suggests that the hydrologically meaningful feature set acts as an implicit regularizer: physically grounded predictors (terrain derivatives, antecedent precipitation, HAND) constrain the model toward plausible inundation patterns even when a quarter of training labels are intentionally mislabeled. These results indicate that the models are robust to moderate levels of ground truth uncertainty (Schumann et al., 2009), lending quantitative rigor to the qualitative assessments in Figure 4 and Figure 9. This finding has practical implications: it suggests that the effective ceiling on model performance may be set more by systematic biases in remote sensing products (e.g., consistent underdetection of small water bodies) than by random label noise, and that physics-constrained feature engineering offers a tractable path forward even in the absence of perfect ground truth. An open question remains regarding how much noise is too much. Exploring higher corruption rates to identify

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

the breaking point of this resilience represents a promising direction for future research. XGBoost Multi Objective generates similarly significant but insubstantial degradations in model performance, indicating that these smoothing operations on the underlying features are not beneficial.

3.3 Basin-Scale Training and Regional Bias (H2)

The cross-basin transfer experiments (Figure 8) show that models trained in one basin generally perform worse when applied to other basins, with 53 out of 56 transfer scenarios significantly underperforming relative to in-basin models. The notable exception is the Uganda basin (Location 5, a lakeside flood on Lake Victoria), where the transfer model performs well due to a more balanced distribution of water and non-water classes. This latter finding emphasizes the importance of representative and balanced training data for generalization across regions (He and Garcia, 2009). Meanwhile, the limited transferability of models across basins is a well-documented challenge, often attributed to regional biases and differences in hydrological regimes. Together, these results reinforce the core rationale of BAGIM: empirical models must be calibrated at the basin scale to account for systemic data biases and blind spots in global datasets, rather than relying on globally uniform training that may propagate regional inconsistencies.

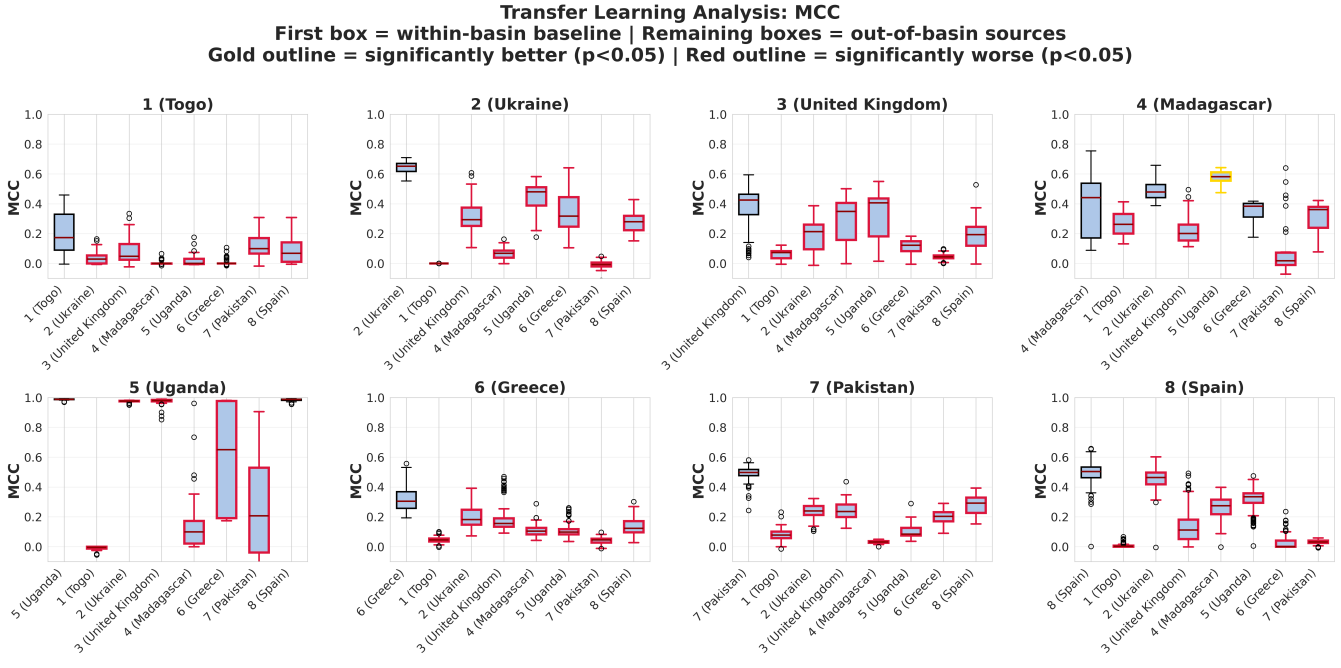

Figure 8. Comparing the performance (MCC - Matthews Correlation Coefficient) of in-basin trained baseline XGBoost models (first boxplot in each subfigure) with XGBoost models trained in other basins and deployed to the basin of interest. Models are highlighted Gold or Red to indicate statistically significant ($p < 0.05$) differences from the in-basin trained baseline XGBoost model, as computed by the Mann-Whitney U test.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

3.4 Generalizing to the Extreme (H3)

The ability of empirical ML models to generalize to rare or extreme flood events that exceed the range of conditions present in the training data remains a central challenge in hydrological prediction and operational flood mapping (Sit et al., 2020; Schoppa et al., 2020; Bentivoglio et al., 2022).

Figure 9 presents an analysis of one such example from the STURM-Flood evaluation set: XGBoost model performance during the 2020-02-08 Madagascar flood (Location 4, a fluvial event), a high-magnitude event that falls entirely outside the temporal range of our OPERA-based training data, with masking derived from S2 imagery with implicit biases that differ from those in OPERA's S1 masks.

The XGBoost model, trained exclusively on non-extreme events, demonstrates a qualified ability to predict increased areal flood extent during this extreme event. When benchmarked against the STURM reference, the model performance declines to an F1 score of 0.28; however, against the coincident Copernicus S2 classification, this improves to an F1 score of 0.45. These results also further highlight the substantial uncertainty in remote sensing-derived flood masks, with STURM predicting the Copernicus S2 mask with an F1 of only 0.56. The severity of this disagreement between flood mapping products is not a unique case, but represents a commonly identified systemic issue in the field (Xu et al., 2025; Afshari et al., 2023; Msabi and Makonyo, 2021) with only a 30-40% level of agreement among different products (Risling et al., 2024).

Notably, the XGBoost model systematically underestimates the full spatial extent of inundation in the central floodplain. This underestimation is visually apparent in the confusion matrix-style maps and is consistent with the conservative bias observed in ML models when extrapolating beyond their training distribution (Gulrajani and Lopez-Paz, 2020; Arjovsky et al., 2019). Yet, the XGBoost model correctly delineates the main river channel and a small northern stream—features that are missed by the STURM mask. The discrepancies between STURM and S2 references, and the model's ability to capture small water bodies missed by both, underscore the limitations of simplistic accuracy-based assessments and the need for hydrologically informed evaluation criteria (Merwade et al., 2008; Datta et al., 2023), while also highlighting the conservative bias in ML-based estimates of extreme out-of-sample events.

Figure 10 extends this analysis to a suite of extreme events across all eight study locations, providing a broader perspective on model generalization. Across these out-of-sample extremes, model performance generally declines but remains within a reasonable range: median accuracy spans 0.65 to 0.95, F1 scores range from 0.5–0.95 for half the locations, around 0.4 for two, and 0.1–0.2 for the most challenging sites; and median ROC-AUC remains above 0.5 across all locations.

A consistent pattern emerges: **precision increases while recall decreases** during extreme events. This indicates that the models become more conservative, producing fewer false positives but missing a greater proportion of the true inundated area. Such a trade-off is well-documented in the literature, where ML models tend to underpredict the spatial extent of rare, high-magnitude floods, prioritizing reliability over completeness in the face of uncertainty (Mosavi et al., 2018; Bentivoglio et al., 2022). Thus, these ML-based predictions of inundation for out-of-sample extremes may be interpreted as a conservative estimate of minimal likely impacts.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

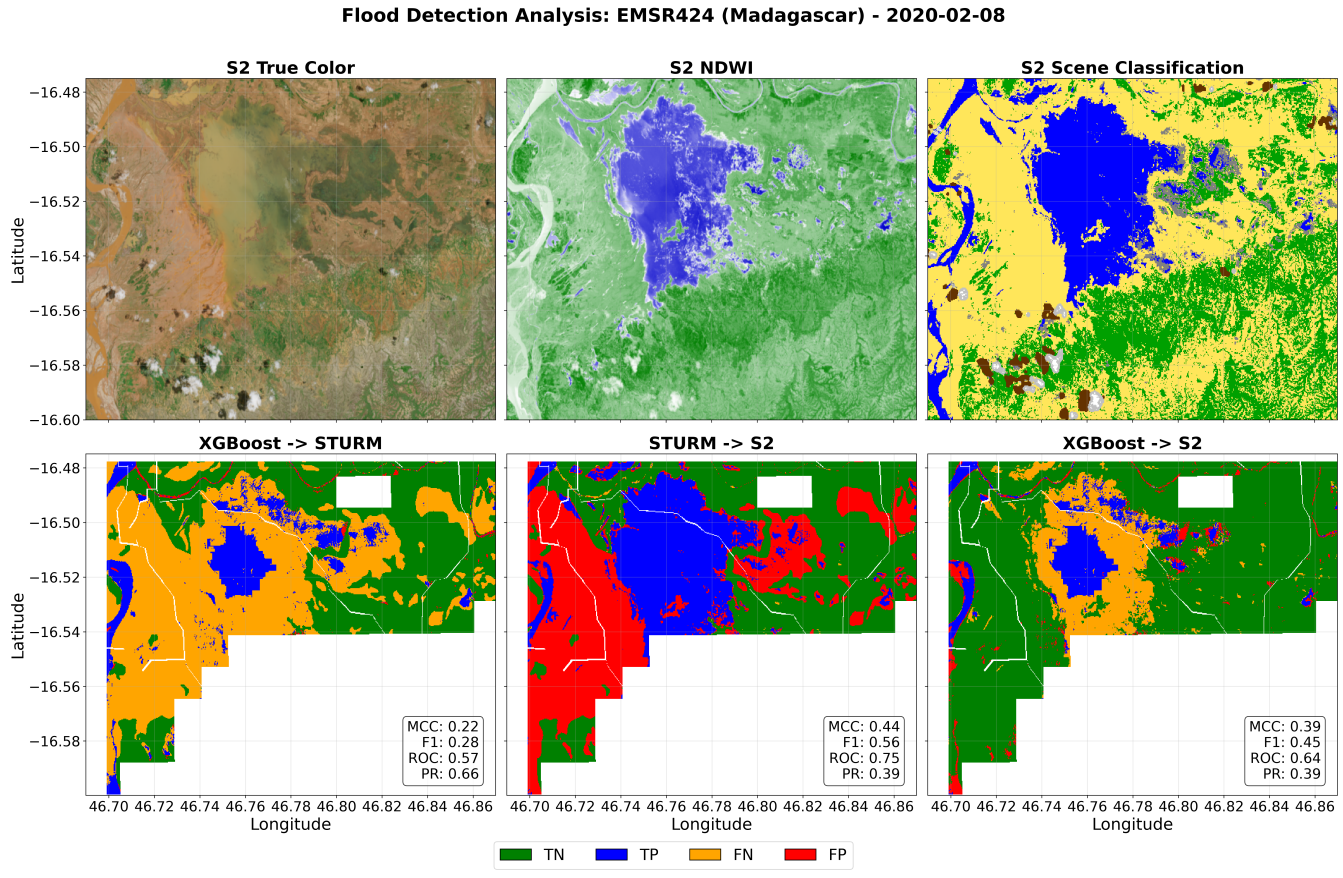

Figure 9. Clockwise from top left: S2 True Color is a true color image of the flooded region (Madagascar) on 2020-02-08. S2 NDWI is the normalized difference water index as computed on the image by the Copernicus online portal. S2 Classification, also derived directly from the Copernicus online portal, indicates water in blue. XGBoost -> S2 presents a "confusion matrix" style map indicating the "accuracy" of our XGBoost model when predicting Copernicus's online portal's water mask. STURM -> S2 presents STURM when predicting Copernicus's online portal's generic classification map. XGBoost -> STURM presents our XGBoost model when predicting STURM. Note that the XGBoost model over indexes on enhanced flooding in the main river corridor, correctly identifies the small stream to the north, and then misses by a large factor the extensive flooding in the central plain. Meanwhile, the S2 mask accurately captures the main central flooded plain but misses the secondary flooded area to the west and portions of the stream. And STURM accurately captures both primary flooded areas in the central plain, but entirely misses the small northern stream and overestimates flooding immediately to the east of the western river.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

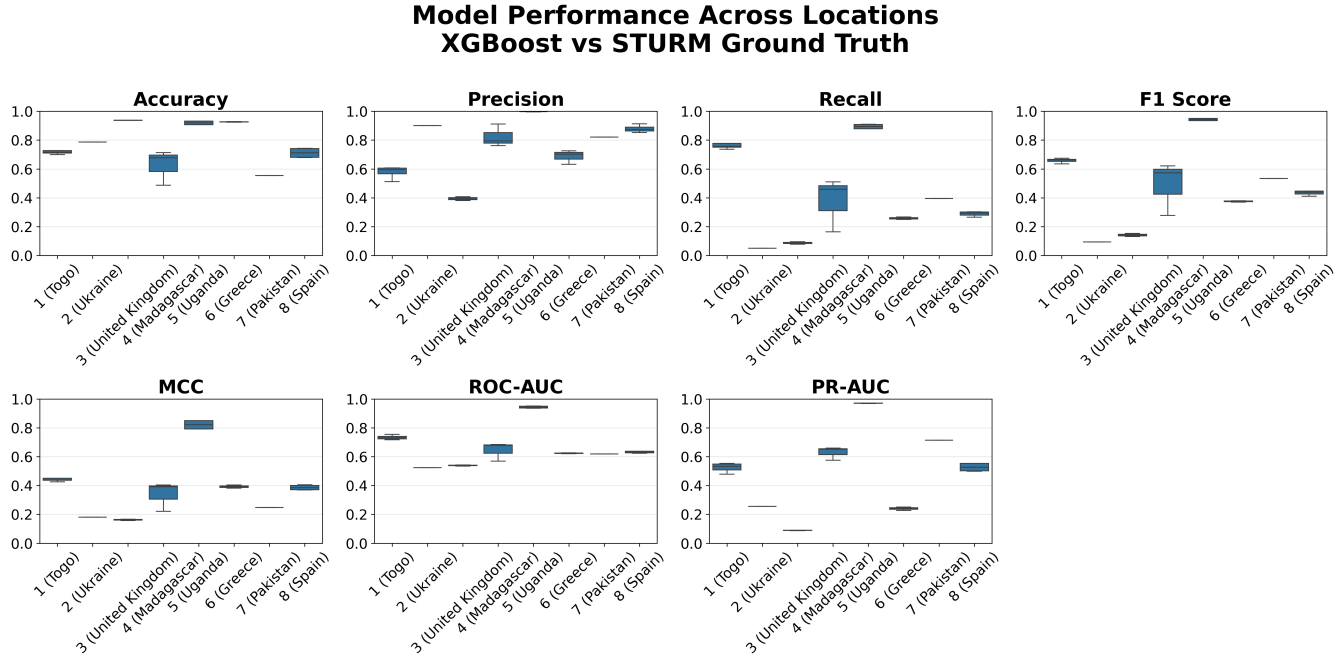

Figure 10. Performance metrics for XGBoost models evaluated on extreme flood events from the STURM-Flood dataset across all eight study locations. Models were trained exclusively on non-extreme events from OPERA DSWx-S1. The consistent pattern of increased precision and decreased recall during extremes indicates conservative bias when extrapolating beyond the training distribution.

The observed trade-off between precision and recall during extremes is consistent with findings from recent reviews, which highlight the difficulty of extrapolating ML models beyond the range of observed training data (Mosavi et al., 2018; Bentivoglio et al., 2022). The need for diverse and representative training data, including extreme cases, is a recurring theme in the literature on flood hazard modeling.

3.5 Ensemble Approaches

Among the four majority-vote ensembles defined in Section 2.4.4, the *Architecture Ensemble* consistently outperforms the XGBoost Baseline in five out of seven evaluation metrics and is never significantly worse (Figure 11). This result aligns with recent literature demonstrating that ensembles of diverse, high-quality models (particularly those spanning different algorithmic families) can effectively reduce overfitting and model-specific biases, leading to improved accuracy and robustness in hydrological prediction tasks (Zounemat-Kermani et al., 2021; Wang et al., 2021; Liu et al., 2021). The *Feature Ensemble*, aggregating models trained on different feature sets and data perturbations, performs similarly to the baseline, suggesting strong consensus among the constituent models and reinforcing the importance of hydrologically meaningful feature engineering (Kratzert et al., 2019; Nearing et al., 2021).

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

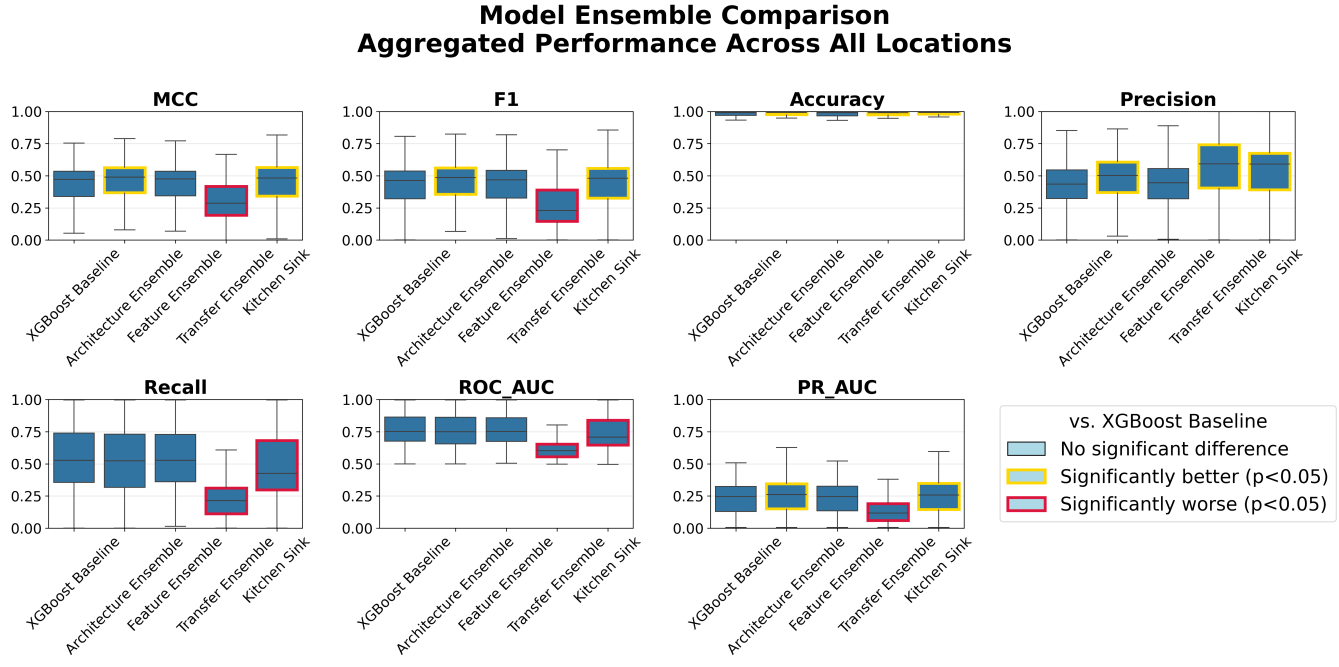

Figure 11. Results of ensembling experiments, comparing the baseline XGBoost model (XGB - All Features) to a variety of model ensembles with pixel-level majority voting. Models are highlighted Gold or Red to indicate statistically significant ($p < 0.05$) differences from the XGB - All Features baseline model, as computed by the Mann-Whitney U test.

In contrast, the *Transfer Ensemble*, which aggregates models trained in different basins, significantly underperforms the baseline in six of seven metrics. This underperformance indicates that transfer models are not merely more imprecise, but are often fundamentally biased or not hydrologically realistic in the target domain. Such findings are consistent with ensemble theory and recent hydrological studies, which caution that the inclusion of poorly adapted or low-quality models can degrade ensemble performance, especially when their errors are systematic rather than random (He and Garcia, 2009; Wang et al., 2021). The so-called *Kitchen Sink* ensemble, which combines all architecture, feature, and transfer models, performs better than the baseline in five metrics but worse in two, further illustrating that indiscriminate inclusion of transfer models can outweigh the benefits of diversity and majority voting.

3.6 Limitations and Path Forward

A central challenge for empirical inundation modeling is the substantial disagreement among remote sensing-derived inundation products, as illustrated in Figure 4 and Figure 9. Different sensors, algorithms, and temporal windows yield systematically different inundation masks, and no single product can be treated as definitive ground truth. This uncertainty ceiling is unlikely to be resolved quickly through improved remote sensing alone.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Our results suggest that physics-constrained machine learning offers a tractable path forward. The resilience of model performance to 25% label noise (Figure 7) demonstrates that hydrologically meaningful feature engineering acts as an implicit regularizer, guiding models toward physically plausible predictions even when trained on imperfect labels. Rather than waiting for perfect ground truth, practitioners can leverage this physics-constrained approach to extract robust signal from noisy observations (Tashie et al., 2021). The 14-month training window utilized here may not capture the full range of seasonal and interannual hydrological variability, and the assumption of stationarity in feature-inundation relationships may limit model reliability as precipitation regimes shift under climate change. Nonetheless, this represents a pragmatic alternative to the paradigm of ever-larger training datasets: careful feature engineering grounded in hydrological first principles can compensate for (and may ultimately matter more than) incremental improvements in label quality.

Complementary strategies may further improve reliability. Multi-product ensemble approaches that synthesize information across sensors and algorithms could reduce the influence of product-specific biases. And targeted validation campaigns in under-monitored regions would help characterize systematic errors in existing products.

4 Conclusions

This study introduced the Basin-Aware Global Inundation Modeling framework (BAGIM) and evaluated three hypotheses regarding the design of scalable, regionally calibrated inundation prediction systems. Our findings support the conclusion that hydrologically meaningful feature engineering is more impactful than architectural complexity for inundation prediction (H1): tree-based ensembles consistently outperformed more complex deep learning architectures while requiring no specialized computational infrastructure, underscoring the importance of model selection and tuning. Feature engineering, particularly the use of derived and smoothed variables, emerged as a key driver of model skill, often compensating for the absence of direct streamflow estimates. Notably, features commonly assumed essential for operational flood forecasting (including coincident river-basin streamflow) were neither sufficient nor strictly necessary for reliable inundation prediction, with well-engineered meteorological and terrain features achieving comparable performance without explicit streamflow inputs.

Cross-basin transfer experiments confirmed that basin-scale training is essential for mitigating regional biases in global geospatial datasets (H2), with the vast majority of transfer scenarios producing statistically significant performance degradation. This underscores a critical finding with broad practical implications: certain topographic and land-cover features may be highly predictive in specific hydroclimatic contexts while contributing little in others, reinforcing the value of basin-aware calibration that can leverage locally relevant predictors. Models trained on non-extreme events demonstrated directionally correct generalization to extreme flood events (H3), though with conservative bias manifested as higher precision but lower recall, a trade-off with important implications for operational deployment where false negatives carry significant consequences. Nonetheless, the directional correctness and hydrologically realistic predictions of extreme inundation events by models trained only on moderate events suggest potential utility in out-of-sample inundation forecasts (Acuña Espinoza et al., 2025).

These findings also point toward promising directions for future work. The integration of probabilistic or ensemble approaches represents a compelling avenue for quantifying prediction uncertainty and further improving the reliability of inunda-

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Hypo.	Tests	Evidence	Findings
H1	Compares six architectures across eight basins; ablated feature sets including basic operational features, no streamflow, multi-objective derived features, and 25% label noise	Fig. 5, 7	**Supported**: XGBoost outperforms deep learning; basic operational features alone are insufficient; streamflow not strictly necessary
H2	Trained per-basin XGBoost models and evaluated all 56 cross-basin transfer scenarios; tested architecture, feature, transfer, and combined ensembles	Fig. 8, 11	**Supported**: 53 of 56 transfer scenarios significantly underperform in-basin models; transfer ensembles degrade performance
H3	Evaluated models trained exclusively on non-extreme OPERA observations against STURM-Flood extreme events across all eight study sites	Fig. 9, 10	**Partially supported**: Directionally correct with increased precision but decreased recall, indicating conservative bias

Table 3. Summary of hypothesis testing results. Each hypothesis is linked to the experimental test designed to evaluate it and the key figures presenting the evidence.

tion estimates across varying conditions. Crucially, BAGIM's modular, basin-scale architecture accommodates such methodological advances incrementally, without requiring the global retraining infrastructure or proprietary data dependencies that constrain many existing approaches.

Together, these results demonstrate that BAGIM combines the scalability of globally available data with the regional sensitivity of basin-scale calibration, delivering meaningful daily-snapshot inundation predictions through an accessible framework. By depending exclusively on open data sources (but allowing the optional inclusion of regionally specific data) and leveraging computationally efficient tree-based methods, this approach offers a flexible pathway for organizations seeking operational inundation mapping capabilities.

Data Availability

The code and data used in this study are available at https://github.com/ClimateAI/basin-aware-global-inundation-modeling-framework.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Author contributions

AT implemented the workflow, developed the experiments and related code, and prepared the initial draft of the manuscript. DF provided continuous support in experiment design, while EK and IG provided continuous support in model implementation and data processing. EG provided data processing support, and CH provided feature engineering guidance. All authors contributed to the discussion section of the final version of the manuscript.

Declaration of competing interests

The contact author has declared that none of the authors has any competing interests.

5 Appendix

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

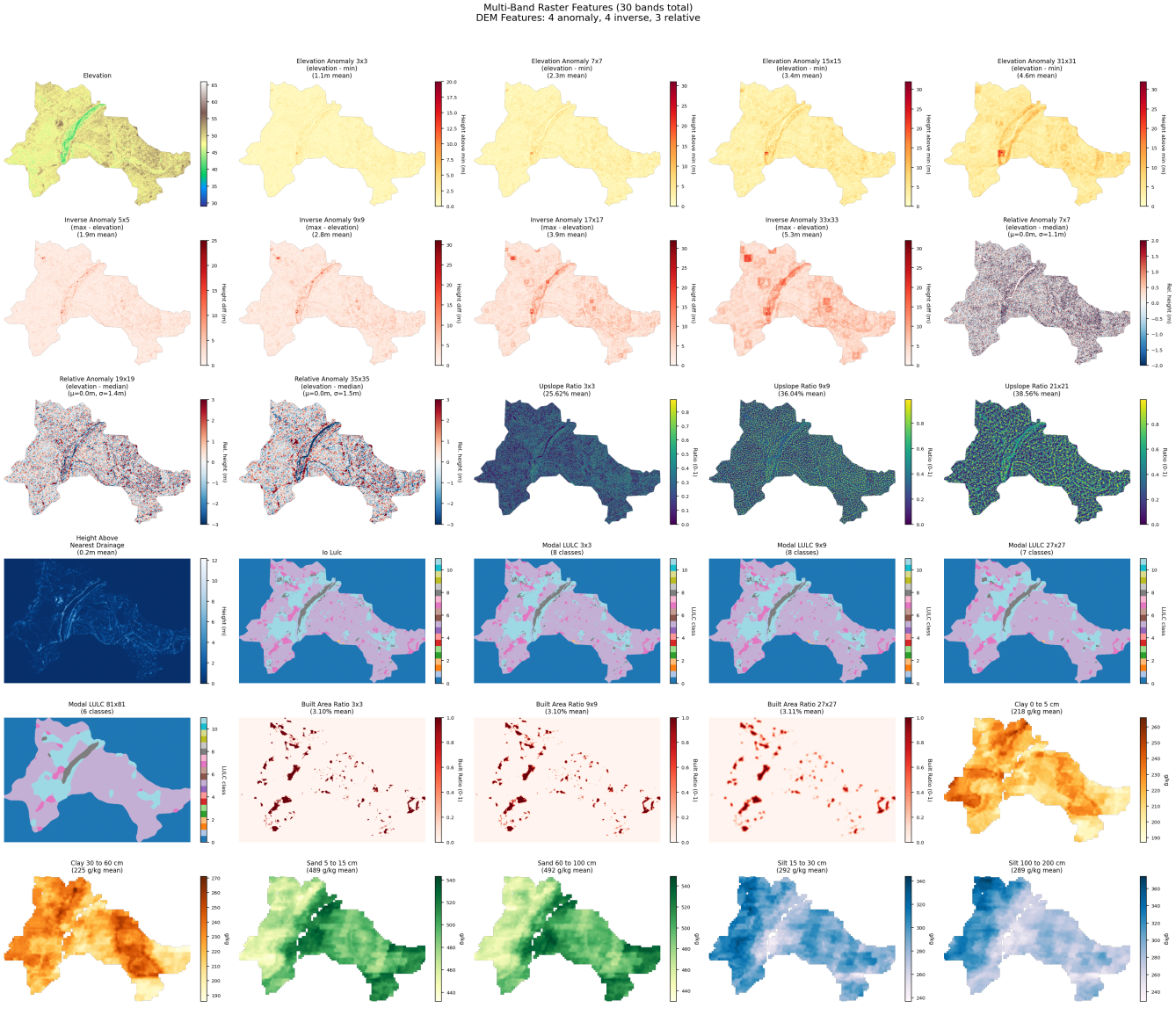

Figure A1. Example (target subwatershed of location 7 (Pakistan)) of all static geospatial datasets and derived features used in model training.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

References

Acharya, T. D., Subedi, A., Huang, H., and Lee, D. H.: Evaluation of Machine Learning Algorithms for Surface Water Extraction in a Landsat 8 Scene of Nepal, Sensors, 19, 2769, https://doi.org/10.3390/s19122769, 2019.

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., and et al.: Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, Hydrology and Earth System Sciences, 29, 1277–1302, https://doi.org/10.5194/hess-29-1277-2025, 2025.

Afshari, S. et al.: Comparative analysis of performance and mechanisms of flood inundation map generation using Height Above Nearest Drainage, Environmental Modelling & Software, 160, 105 526, https://doi.org/10.1016/j.envsoft.2022.105526, 2023.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, 2019.

Alfieri, L., Feyen, L., and Salamon, P.: GloFAS: Global Flood Awareness System, ECMWF Newsletter, 137, 17–22, 2013.

Arad, B., Matias, Y., et al.: Flood forecasting with machine learning models in an operational framework, Hydrology and Earth System Sciences, 26, 1831–1847, https://doi.org/10.5194/hess-26-1831-2022, 2022.

Arik, S. O. and Pfister, T.: TabNet: Attentive Interpretable Tabular Learning, Proceedings of the AAAI Conference on Artificial Intelligence, 35, 6679–6687, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D.: Invariant Risk Minimization, arXiv preprint arXiv:1907.02893, 2019.

Bentivoglio, D., Dottori, F., Alfieri, L., and et al.: Deep learning for flood mapping: A review, Remote Sensing, 14, 567, https://doi.org/10.3390/rs14030567, 2022.

Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrological processes, 6, 279–298, 1992.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Cao, H., Tian, Y., Liu, Y., and Wang, R.: Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion, Scientific Reports, 14, 16 132, https://doi.org/10.1038/s41598-024-67113-7, 2024.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357, 2002.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chicco, D. and Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC genomics, 21, 6, 2020.

Crippen, R., Buckley, S., Agram, P., Belz, E., Gurrola, E., Hensley, S., Kobrick, M., Lavalle, M., Martin, J., Neumann, M., Nguyen, Q., Rosen, P., Shimada, J., Simard, M., and Tung, W.: NASADEM GLOBAL ELEVATION MODEL: METHODS AND PROGRESS, in: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLI-B4, pp. 125–128, https://doi.org/10.5194/isprs-archives-XLI-B4-125-2016, 2016.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F.: Scientific machine learning through physics–informed neural networks: Where we are and what's next, Journal of Scientific Computing, 92, 88, 2022.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Datta, S., Nawaz, S., Hossen, M. N., Karim, M. E., Juthy, N. T., Hossain, M. L., and Kabir, M. H.: Flood risk assessment in developing countries: Dealing with data quality and availability, in: Handbook of flood risk management in developing countries, pp. 197–216, Routledge, 2023.

Falcon, W. A.: Pytorch lightning, GitHub, 3, 2019.

Ferreira, V. G., Asiah, Z., Xu, J., Gong, Z., and Andam-Akorful, S. A.: Land Water-Storage Variability Over West Africa: Inferences From Space-Borne Sensors, https://doi.org/10.3390/w10040380, 2018.

Frame, J. M., Nearing, G. S., Kratzert, F., and Rahmani, V.: Post-processing the US National Water Model with a Long Short-Term Memory network, Environmental Research Letters, 15, 104 020, https://doi.org/10.1088/1748-9326/aba927, 2020.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling & Software, 135, 104 926, https://doi.org/10.1016/j.envsoft.2020.104926, 2021.

Gnann, S., Baldwin, J. W., Cuthbert, M. O., Gleeson, T., Schwanghart, W., and Wagener, T.: The Influence of Topography on the Global Terrestrial Water Cycle, Reviews of Geophysics, 63, e2023RG000 810, https://doi.org/10.1029/2023RG000810, 2025.

Gulrajani, I. and Lopez-Paz, D.: In Search of Lost Domain Generalization, in: International Conference on Learning Representations (ICLR), 2020.

He, H. and Garcia, E. A.: Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering, 21, 1263–1284, https://doi.org/10.1109/TKDE.2008.239, 2009.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., and ... Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hub, G. F.: Flood Hub: Real-time Flood Forecasting, https://floodhub.google.com/, accessed: 2025-12-10, 2022.

Inman, V. L. and Lyons, M.: Automated Inundation Mapping Over Large Areas Using Landsat Data and Google Earth Engine, https://doi.org/10.3390/rs12081348, 2020.

Jafarzadegan, K. and et al.: Recent advances and new frontiers in riverine and coastal flood modeling, Reviews of Geophysics, 61, e2022RG000 788, https://doi.org/10.1029/2022RG000788, 2023.

Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., and Pender, G.: A deep convolutional neural network model for rapid prediction of fluvial flood inundation, Journal of Hydrology, 590, 125 481, https://doi.org/10.1016/j.jhydrol.2020.125481, 2020.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S.: Global land use/land cover with Sentinel 2 and deep learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4704–4707, IEEE, https://doi.org/10.1109/IGARSS47720.2021.9553495, 2021.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980, 2014.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, W03S04, https://doi.org/10.1029/2005WR004362, 2006.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Kratzert, F., Klotz, D., Herrnegger, M., and et al.: 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, 2019.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. S.: Toward learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 25, 3245–3261, https://doi.org/10.5194/hess-25-3245-2021, 2021.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., et al.: Caravan-A global community dataset for large-sample hydrology, Scientific Data, 10, 61, 2023.

Laboratory, N. J. P.: OPERA Dynamic Surface Water Extent (DSWx) Product, Available at https://opera.jpl.nasa.gov/, accessed: 2025-12-10, 2023.

Leopold, L. B.: Hydrology for urban land planning: A guidebook on the hydrologic effects of urban land use, vol. 554, US Geological Survey, 1968.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Scientific Data, 6, 283, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Liu, Y., Wang, Y., Zhang, Y., et al.: Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size, Journal of Hydrology, 603, 127 553, https://doi.org/10.1016/j.jhydrol.2021.127553, 2021.

Llugsi, R., El Yacoubi, S., Fontaine, A., and Lupera, P.: Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito, in: 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), pp. 1–6, IEEE, 2021.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Mangukiya, N., Kushwaha, S., and Sharma, A.: A novel multi-model ensemble framework for fluvial flood inundation mapping, Environmental Modelling and Software, 173, 106 163, https://doi.org/10.1016/j.envsoft.2024.106163, 2024.

Martinis, S., Groth, S., Wieland, M., Knopp, L., and Rättich, M.: Towards a global seasonal and permanent reference water product from Sentinel-1/2 data for improved flood mapping, Remote Sensing of Environment, 278, 113 077, https://doi.org/10.1016/j.rse.2022.113077, 2022.

Merwade, V., Olivera, F., Arabi, M., and Edleman, S.: Uncertainty in flood inundation mapping: Current issues and future directions, Journal of hydrologic engineering, 13, 608–620, 2008.

Milletari, F., Navab, N., and Ahmadi, S.-A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571, IEEE, 2016.

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

Msabi, M. M. and Makonyo, M.: Flood hazard mapping methods: A review, Journal of Hydrology, 603, 127 401, https://doi.org/10.1016/j.jhydrol.2021.127401, 2021.

Neal, J., Villanueva, I., Wright, N., Willis, T., Fewtrell, T., and Bates, P.: How much physical complexity is needed to model flood inundation?, Hydrological Processes, 26, 2264–2282, 2012.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Nelder, J. A. and Wedderburn, R. W. M.: Generalized Linear Models, Journal of the Royal Statistical Society: Series A (General), 135, 370–384, 1972.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo, M., and Saleska, S.: Height Above the Nearest Drainage – a hydrologically relevant new terrain model, Journal of Hydrology, 404, 13–29, https://doi.org/10.1016/j.jhydrol.2011.03.051, 2011.

Notarangelo, N., Wirion, C., and van Winsen, F.: STURM-Flood: a curated dataset for deep learning-based flood extent mapping leveraging Sentinel-1 and Sentinel-2 imagery, Big Earth Data, pp. 1–27, 2025.

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, 418–422, https://doi.org/10.1038/nature20584, 2016.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, Soil, 7, 217–240, 2021.

Powers, D. M. W.: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation, Journal of Machine Learning Technologies, 2, 37–63, 2011.

Risling, A., Lindersson, S., and Brandimarte, L.: A comparison of global flood models using Sentinel-1 and a change detection approach, Natural Hazards, 120, 11 133–11 152, https://doi.org/10.1007/s11069-024-06629-7, 2024.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241, Springer, Cham, 2015.

Rosser, J. F., Leibovici, D. G., and Jackson, M. J.: Rapid flood inundation mapping using social media, remote sensing and topographic data, Natural Hazards, 87, 103–120, https://doi.org/10.1007/s11069-017-2755-0, 2017.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323, 533–536, 1986.

Sajjad, A., Lu, J., Chen, X., Chisenga, C., and Mazhar, N.: Rapid Assessment of Riverine Flood Inundation in Chenab Floodplain Using Remote Sensing Techniques, https://doi.org/10.1186/s40677-023-00236-7, 2023.

Schoppa, L., Disse, M., and Bachmair, S.: Evaluating the performance of random forest for large-scale flood discharge simulation, Journal of Hydrology, 590, 125 423, https://doi.org/10.1016/j.jhydrol.2020.125423, 2020.

Schumann, G. J.-P., Di Baldassarre, G., Alsdorf, D. E., and Bates, P. D.: The utility of spaceborne radar to render flood inundation maps based on multialgorithm ensembles, IEEE Transactions on Geoscience and Remote Sensing, 47, 2801–2807, https://doi.org/10.1109/TGRS.2009.2017937, 2009.

Sharma, N. and Saharia, M.: DeepSARFlood: Rapid and automated SAR-based flood inundation mapping using vision transformer-based deep ensembles with uncertainty estimates, Science of Remote Sensing, 13, 100 203, https://doi.org/10.1016/j.srs.2025.100203, 2025.

Sheffield, J., Wood, E. F., and Roderick, M. L.: Drought: Past problems and future scenarios, Earth-Science Reviews, 105, 37–54, https://doi.org/10.1016/j.earscirev.2011.12.002, 2012.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I.: A comprehensive review of deep learning applications in hydrology and water resources, Water Science and Technology, 82, 2635–2670, https://doi.org/10.2166/wst.2020.369, 2020.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Sivakumar, B. and Singh, V. P.: Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework, Hydrology and Earth System Sciences, 16, 4119–4131, https://doi.org/10.5194/hess-16-4119-2012, 2012.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian optimization of machine learning algorithms, Advances in Neural Information Processing Systems, 25, 2951–2959, 2012.

Tashie, A., Pavelsky, T., Band, L., and Topp, S.: Watershed-scale effective hydraulic properties of the continental United States, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 440, 2021.

Tashie, A., Pavelsky, T., and Kumar, M.: A Calibration-free groundwater module for improving predictions of low flows, Water Resources Research, 58, e2021WR030 800, 2022.

Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F., Dutta, D., and Kim, S.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, Environmental modelling & software, 90, 201–216, 2017.

Tharme, R. E.: A Global Perspective on Environmental Flow Assessment: Emerging Trends in the Development and Application of Environmental Flow Methodologies for Rivers, https://doi.org/10.1002/rra.736, 2003.

Trochim, E., Prakash, A., Kane, D. L., and Romanovsky, V. E.: Remote Sensing of Water Tracks, https://doi.org/10.1002/2015ea000112, 2016.

Uday, G., Purse, B. V., Kelley, D. I., Vanak, A., Samrat, A., Chaudhary, A., Rahman, M., and Gerard, F. F.: Radar versus optical: The impact of cloud cover when mapping seasonal surface water for health applications in monsoon-affected India, PLOS ONE, 20, e0314 033, https://doi.org/10.1371/journal.pone.0314033, 2025.

Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T., and Singh, G.: Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover, Remote Sensing, 14, 4101, https://doi.org/10.3390/rs14164101, 2022.

Wang, C., Pavelsky, T. M., Yao, F., Yang, X., Zhang, S., Chapman, B., Song, C., Sebastian, A., Frizzelle, B. G., Frankenberg, E., and Clinton, N.: Flood Extent Mapping During Hurricane Florence With Repeat-Pass L-Band UAVSAR Images, https://doi.org/10.1029/2021wr030606, 2022.

Wang, J., Wang, G., Elmahdi, A., Bao, Z., Yang, Q., Shu, Z., and Song, M.: Comparison of hydrological model ensemble forecasting based on multiple members and ensemble methods, Open Geosciences, 13, 401–415, https://doi.org/10.1515/geo-2020-0239, 2021.

Wechsler, S. P.: Perceptions of Digital Elevation Model Uncertainty by DEM Users, URISA Journal, 15, 57–64, 2003.

Xu, Q., Shi, Y., Zhao, J., and Zhu, X. X.: FloodCastBench: A large-scale dataset and foundation models for flood modeling and forecasting, Scientific Data, 12, 431, 2025.

Yamazaki, D., Kanae, S., Kim, H., and Oki, T.: A Physically Based Description of Floodplain Inundation Dynamics in a Global River Routing Model, https://doi.org/10.1029/2010wr009726, 2011.

Yamazaki, D., O'Loughlin, F., Trigg, M., et al.: Comparative analysis of global DEMs for hydrological applications in Africa: SRTM, ASTER GDEM, and TanDEM-X, Hydrology and Earth System Sciences, 24, 6059–6078, https://doi.org/10.5194/hess-24-6059-2020, 2020.

Yan, Y., Chen, M., Shyu, M.-L., and Chen, S.-C.: Deep learning for imbalanced multimedia data classification, in: 2015 IEEE international symposium on multimedia (ISM), pp. 483–488, IEEE, 2015.

Yang, X., Qin, Q., Grussenmeyer, P., and Koehl, M.: Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery, Remote Sensing of Environment, 219, 259–270, https://doi.org/10.1016/j.rse.2018.10.046, 2018.

Zhang, B., Ouyang, C., Cui, P., Xu, Q., Wang, D., Zhang, F., Li, Z., Fan, L., Lovati, M., Liu, Y., and Zhang, Q.: Deep learning for cross-region streamflow and flood forecasting at a global scale, The Innovation, 5, 100 617, https://doi.org/10.1016/j.xinn.2024.100617, 2024.

https://doi.org/10.5194/egusphere-2026-1527
Preprint. Discussion started: 24 April 2026

Zhao, G., Pang, B., Xu, Z., Cui, L., Wang, J., Zuo, D., and Peng, D.: Improving urban flood susceptibility mapping using transfer learning, Journal of Hydrology, 603, 126 777, https://doi.org/10.1016/j.jhydrol.2021.126777, 2021.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, Journal of Hydrology, 598, 126 266, https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.

780 Zsoter, E., Salamon, P., Prudhomme, C., et al.: Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data, Environmental Modelling & Software, 124, 104 578, https://doi.org/10.1016/j.envsoft.2019.104578, 2020.