

Mapping maximum floodwater footprints from sustained flood traces

Yiling Lin^{a,b}, Xie Hu^{a,b,*}, Xiekang Wang^c, Yunxia Zhang^d

^a College of Urban and Environmental Sciences, State Key laboratory of Vegetation Structure, Function and Construction (VegLab), and Key Laboratory for Earth Surface Processes of the Ministry of Education, Peking University, Beijing, China

^b Beijing Key Laboratory of Spatio-temporal Perception and Urban Resilience, Beijing, China

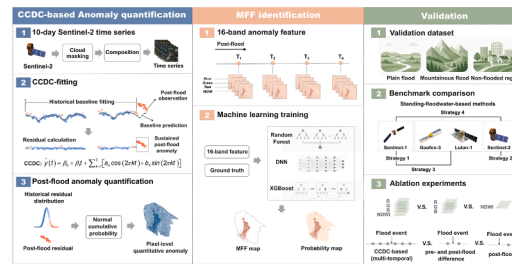
^c State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu, China

^d National Disaster Reduction Center of China, Ministry of Emergency Management, Beijing, China

HIGHLIGHTS

- Flood traces recover maximum flood footprints missed between satellite revisits.
- Post-flood NDWI anomalies are reliable indicators for mapping flood traces.
- Historical baselines improve flood mapping by filtering seasonal and spatial noise.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Maximum floodwater footprints (MFF)
Flood mapping
NDWI
SAR

ABSTRACT

Remote-sensing-based flood extent mapping is essential for flood management. However, existing approaches that target standing floodwater often underestimate Maximum Flood Footprints (MFF) because inundation can expand and recede between satellite revisits. Here, we address this limitation by developing a flood-trace-based MFF extraction framework that uses sustained post-flood traces as delayed evidence of maximum inundation. To make flood traces reliably detectable, the framework introduces a trace-oriented anomaly detection strategy that uses Normalized Difference Water Index (NDWI) anomalies as the primary indicator and CCDC-based historical baselines to separate flood-related signals from normal seasonal variability and spatial heterogeneity. Validation on the once-in-a-century 2023 Hai River Basin (HRB) flood event shows that the framework achieves an average F1-score of 81.45%, whereas four widely used standing-floodwater-based approaches attain F1-scores of only 28.24% to 36.41%. At the catchment scale, the proposed method reduces MFF underestimation caused by missed flood-peak observations. Compared with our estimate, MFF derived from Sentinel-1 alone, Sentinel-2 alone, multi-source SAR imagery (Sentinel-1 + Gaofen-3 + Lutan-1), and fusion of Sentinel-1 and Sentinel-2 is 43.3%, 48.0%, 34.2%, and 20.2% smaller, respectively. This work presents an MFF extraction approach that does not depend on the coincidence between satellite acquisitions and flood peaks. Collectively, these findings contribute to more complete flood mapping and underscore the need to account for systematic underestimation in existing flood-extent datasets.

* Corresponding author.

E-mail address: hu.xie@pku.edu.cn (X. Hu).

<https://doi.org/10.1016/j.watres.2026.126783>

Received 20 June 2026; Received in revised form 15 August 2026; Accepted 23 August 2026

Available online 26 August 2026

0043-1354/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Floods are the most widespread natural hazard in terms of affected population and event frequency, while ranking second in economic losses and fourth in total fatalities (Yin et al., 2023; David and Frasson, 2023). The frequency and magnitude of flood events have been escalating in recent years due to global climate change (Gu et al., 2017; Zhang et al., 2015). Therefore, considerable efforts have been devoted to mitigating flood impacts, and accurate flood-extent mapping is fundamental to many of these initiatives: Real-time flood extent mapping enables rapid emergency response, while post-event reconstructions provide essential data for flood impact assessments (Vousdoukas et al., 2026; Han et al., 2024) and flood management strategy evaluation (Nemnem et al., 2025).

With regular revisit intervals and broad spatial coverage, remote sensing has become an essential tool for flood extent mapping in recent years. Optical and Synthetic Aperture Radar (SAR) imagery are the most widely used data for flood mapping (Portalés-Julià et al., 2025). Optical imagery offers long-term observations, and is well suited for flood mapping over decadal timescales (e.g., Han et al., 2024; Tellman et al., 2021; Pekel et al., 2016). In recent years, SAR images become prevailing due to its all-weather, day and night imaging capacities (e.g., McCormack et al., 2022; Misra et al., 2025). Relying on these data, a wide range of flood mapping methods has been developed, ranging from thresholding on SAR backscatter (Martinis et al., 2015; Landuyt et al., 2019) and optically derived Spectral indices (Gao, 1996; Singh et al., 2015) to more advanced deep-learning models. Threshold-based mapping remains widely used in emergency response and large-scale applications due to its efficiency and ease of implementation (e.g., Han et al., 2024; Tellman et al., 2021), whereas deep learning has gained prominence for its improved performance over other approaches (Bereczky et al., 2022; Wu et al., 2023). Despite these advances, deep learning-based methods can still produce substantial errors when applied across large spatial extents and long time periods. Accordingly, considerable efforts have been devoted to reducing these errors through multi-source data fusion and improved model architectures (Qin et al., 2025; Du et al., 2026; Roth et al., 2025; Portalés-Julià et al., 2025).

However, beyond algorithm-induced biases, a more fundamental limitation inherent to remote sensing data is often overlooked: current mapping relies on floodwater observed at satellite acquisition times, but satellite observations are temporally discrete, and flooding that grows and recedes between revisits cannot be captured. Consequently, unless observations coincide with peak inundation, some flooded areas can be missed, leading to systematic underestimation of flood extent. Unfortunately, such coincidence is rarely achieved with current remote-sensing constellations: For rainfall-driven floods, optical observations are often unusable during flood peak stage due to cloud obstruction. Although SAR is less affected by clouds, the relatively long revisit intervals of publicly available SAR data make it less likely to coincide with flood peak. This limitation is even more evident for short-lived, localized events such as flash floods in hilly regions, whose entire evolution can occur between satellite acquisitions. Currently, the proportion of flood events that are entirely missed by remote sensing observations has been quantified: Wagner et al. (2026) reported that Sentinel-1 imagery entirely missed 31 out of 104 (29.8%) major flood events worldwide during 2022–2024. Tarpanelli et al. (2022) showed that, in Europe, even under idealized conditions (two-satellite constellation for each mission and considering both ascending and descending orbits), 42% of flood events would still be completely missed by Sentinel-1, while the fraction increased to 72% for Sentinel-2 due to cloud contamination.

Integrating multi-source remote sensing data can alleviate the underestimation. The combination of optical and SAR imagery is a common practice, exemplified by the joint use of Sentinel-1 and Sentinel-2 (e.g., Fichtner et al., 2023; Chen & Zhao, 2022; Hamidi et al., 2023). Their complementary sensing capabilities can reduce cloud-induced observation gaps (Ahmadi et al., 2025, 2026), while incorporating

terrain information can provide additional topographic constraints to further mitigate Maximum Flood Footprints (MFF) underestimation in areas with incomplete observations (Li et al., 2025). Rapidly developed commercial SAR constellations also provide a critical complement (e.g., Lin et al., 2025; Lan & Wang, 2025). Nevertheless, multi-source integration can only densify temporal sampling and increase the likelihood that observations coincide with peak inundation. As a result, integrating multi-source data does not fundamentally resolve the problem, recovering the full flood-affected area still hinges on whether acquisitions happen to capture the flood peak.

Compared with standing floodwater, flood traces are more durable and can remain detectable for weeks to months after the flood peak. The persistence of such flood traces has been documented in previous studies, including sediment deposition (Chen et al., 2025; Baynes et al., 2023), vegetation stripping (Xu et al., 2025), and crop damage (Rahman et al., 2021; Garcia-Vila et al., 2025). Therefore, when satellite observations are unavailable during peak inundation, sustained flood traces may provide a more practical post-flood indicator for inferring MFF than short-lived floodwater.

Existing studies have exploited different types of flood traces to reconstruct flood-affected areas. Sediment deposition, for example, can preserve evidence of previous flooding, and freshly deposited materials may form relatively smooth surfaces with low radar backscatter, allowing them to be identified from SAR imagery (Ma et al., 2024; Normand & Heggy, 2024). Flood-induced vegetation damage serves as another important type of flood trace. Anomalies in vegetation indices have been used to identify flood-damaged vegetation and infer affected areas, particularly for short-lived flash floods that may be missed by satellite overpasses (e.g., Xu et al., 2025; Bofana et al., 2022). However, these approaches generally target a specific type of flood trace. Because flooding produces different surface responses across land-cover types, reliance on a single trace type may underestimate MFF.

A further limitation is that many flood-trace-based approaches rely on simple bi-temporal pre- and post-flood differencing, which can be confounded by normal land-surface dynamics unrelated to flooding. For example, changes in vegetation indices may reflect normal crop phenology rather than flood-induced damage, making it difficult to isolate flood signals from natural temporal variability (Chen et al., 2019). Similarly, in hydrologically dynamic environments, seasonal fluctuations in wetlands, ponds, and other temporary water bodies may be misclassified as flooding, reducing the reliability of flood maps and necessitating additional exclusion procedures (Lin et al., 2025; Martinis et al., 2022).

Recent advances in disaster change detection demonstrate that historical remote-sensing time series can help distinguish disaster-induced anomalies from normal surface dynamics, offering a promising means of reducing such ambiguity in flood-trace-based MFF reconstruction. For instance, Deijns et al. (2024) use the peak cumulative difference from the mean across multiple spectral-index time series derived from Sentinel-2 observations to detect flash-flood and landslide events. Li et al. (2026) propose AnomalyCD, which leverages historical observations to distinguish disaster-induced anomalies from normal changes associated with human activities. Meanwhile, the rapid development of cloud-computing platforms such as Google Earth Engine (GEE) provides new opportunities for large-scale disaster change detection by enabling efficient processing of long-term satellite time series (Gorelick et al., 2017). For example, DeVries et al. (2020) use temporal anomalies derived from dense Sentinel-1 time series together with historical Landsat water information on GEE to distinguish newly inundated areas from permanent or seasonally varying surface water. Among established time-series change-detection approaches, Continuous Change Detection and Classification (CCDC) is particularly effective at distinguishing anomalous changes from normal seasonal variability. The algorithm continuously models pixel-level seasonality, trends, and breakpoints from satellite time series and was originally developed for land-cover change detection (Zhu & Woodcock, 2014). Its applications have

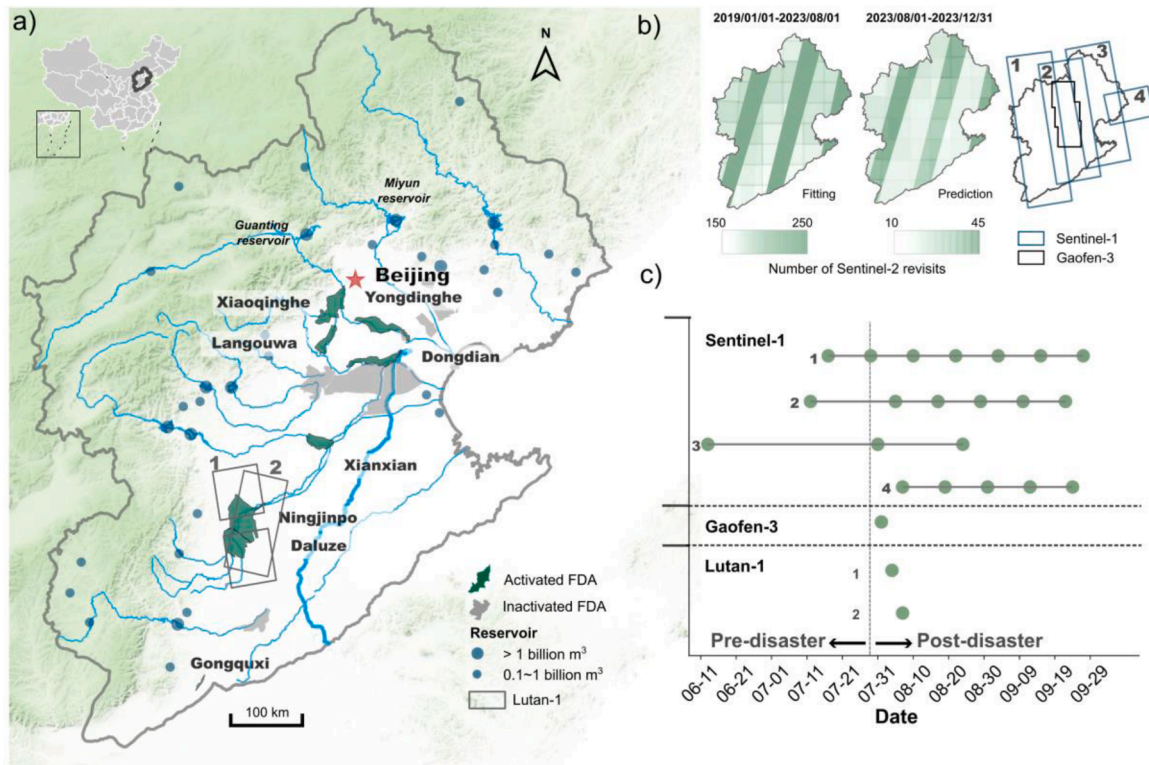


Fig. 1. Study areas and data. (a) Geographic location of the Hai River Basin and the eight flood detention areas activated during the 2023 HRB flood event. Lutan-1 coverage is outlined by grey boxes, with the above labels indicating the corresponding orbit numbers in (c). (b) Sentinel-2 revisit counts for CCDC fitting and post-flood anomaly quantification, with Sentinel-1 and Gaofen-3 acquisition footprints. Labels indicate the corresponding orbit numbers shown in (c). (c) Acquisition dates of the multi-source SAR imagery.

expanded to the detection of specific disturbance processes, including fire, logging, pest damage, and insect defoliation (e.g., Zhang et al., 2022; Feng et al., 2026). The employment of CCDC on Google Earth Engine (GEE) further facilitates its application to large-scale disturbance monitoring. For example, Aufaristama et al. (2024) implement CCDC on GEE to identify the timing and location of landslide disturbances from two decades of Landsat observations.

Despite these advances, the use of CCDC-derived historical baselines for flood-trace detection and the development of appropriate anomaly metrics tailored to flood traces remain unexplored. Moreover, diverse flood-induced surface responses have yet to be systematically integrated within a unified framework for MFF reconstruction. These gaps limit the reliability of flood-trace-based MFF mapping. To address these limitations, we develop an MFF reconstruction framework that integrates multiple types of flood traces and quantifies their anomalies using a probabilistic anomaly metric. The contributions of this framework are threefold. First, it integrates multiple types of flood traces within a unified framework, improving the completeness of MFF mapping across heterogeneous land-surface conditions. Second, it uses post-flood NDWI anomaly as an effective indicator for identifying sustained flood traces, extending use of NDWI from conventional standing-water mapping to flood footprint mapping. Third, it introduces a CCDC-based pixel-level probabilistic anomaly metric tailored to flood trace persistence. By reducing the effects of seasonal cycles and land-surface heterogeneity, this strategy produces reliable anomaly features for machine-learning-based classification, thereby improving the accuracy and robustness of MFF mapping.

The remainder of the paper is organized as follows. Section 2 describes the study area and datasets. Section 3 introduces the flood-trace-based framework for MFF extraction. Section 4 presents the validation results of the MFF extraction framework at both validation sites and the catchment scale. Section 5 provides extended analyses of the proposed

framework and uncertainty estimation. Section 6 concludes the paper.

2. Study area and data

2.1. Study area

We evaluate the model's effectiveness in the 2023 once-in-a-century flood in HRB. Triggered by ~ 4.9 million m^3 of precipitation delivered by Typhoon Doksuri between 28 July and 2 August, the flood event swept across the entire HRB—a basin encompassing the Beijing-Tianjin-Hebei metropolis and covering $\sim 3.3\%$ of China's land area. To temporarily store excess floodwaters and alleviate pressure to downstream areas, eight Flood Detention Areas (FDA) were activated. The scale of the event makes it an ideal natural laboratory to evaluate our method across diverse flood-evolution modes and surface conditions: flooding in the western mountains was short-lived, often receding within a day due to steep terrain, whereas inundation in the eastern plains was extensive and sustained. For instance, in the Dongdian FDA, extensive inundation persisted for nearly a month. The flood also spanned heterogeneous land covers, including farmland, forest, shrubland, and impervious surfaces (Fig. 1a).

2.2. Data

2.2.1. Remote sensing dataset

Multi-temporal Sentinel-2 optical imagery covering the entire HRB was retrieved from Google Earth Engine (GEE) for MFF extraction. Sentinel-2 scenes with more than 70% cloud cover are first excluded, after which clouds, cloud shadows, cirrus, snow/ice, and invalid pixels are masked using the Scene Classification Layer (SCL). We use acquisitions from January 1, 2019 to August 1, 2023, prior to the onset of the 2023 HRB flood, for normal seasonal trajectory modeling. Scenes

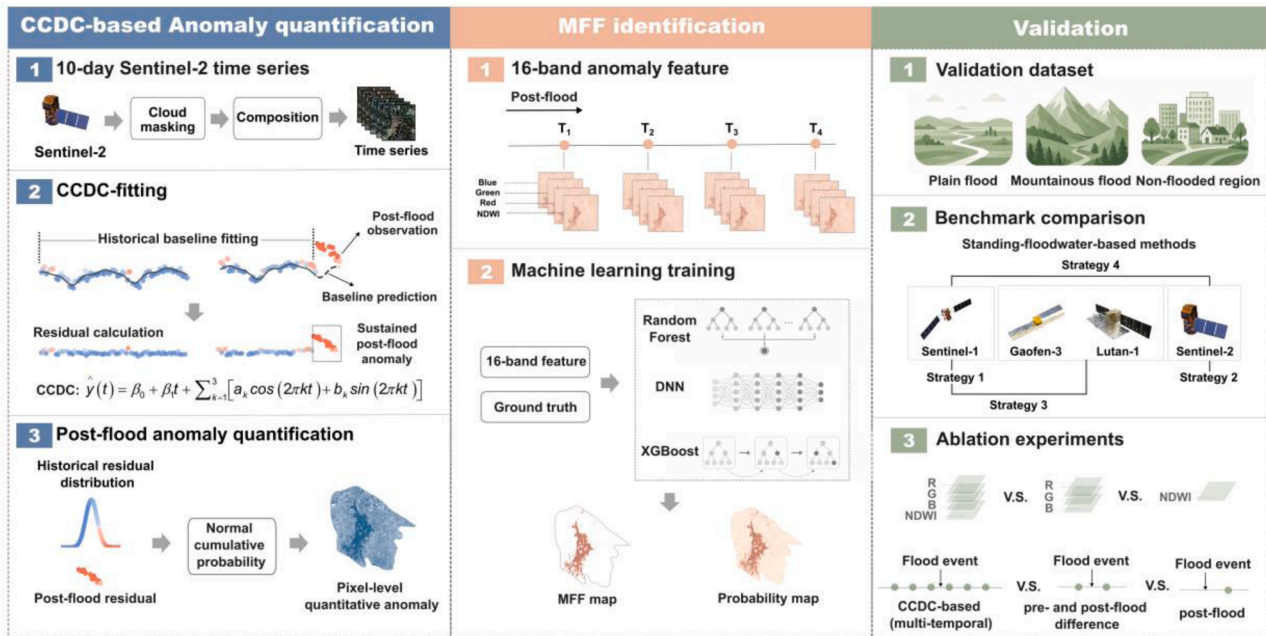


Fig. 2. Workflow of our flood-trace guided framework.

acquired from August 1 to December 31, 2023 are then used for post-flood anomaly detection. We group all available Sentinel-2 observations into 10-day time windows and calculate the mean value within each window to construct a 10-day regular time series. The step converts irregular satellite observations into a regular temporal sequence, ensuring that multi-temporal post-flood anomaly features from different pixels correspond to consistent time windows. Meanwhile, acquisitions from June 1 to December 31, 2021 are excluded to avoid contamination from post-flood anomalies associated with the 2021 HRB flood event (Zhang et al., 2021).

To compare our approach against commonly used flood extent mapping methods targeting standing floodwaters, we collect multi-source SAR images comprising Sentinel-1, Gaofen-3, and Lutan-1 imagery. Multi-temporal Sentinel-1 Radiometric Terrain Correction (RTC) scenes covering the entire HRB are used to implement single-SAR-source flood-mapping method. Pre-flood acquisitions are used to delineate regular water bodies, while the post-flood time series are processed to derive multi-temporal flood extents (Fig. 1b and c). To construct methods combining commercial SAR sensors with Sentinel-1, we further incorporate Gaofen-3 imagery acquired on 1 August over the central HRB, and Lutan-1 imagery acquired on 4 and 7 August over the severely flooded Daluze and Ningjinpo FDA (Fig. 1a and c). In addition, Sentinel-2 scenes from 1-15 August are supplemented, enabling an optical-imagery-based flood mapping.

2.2.2. MFF reference labels

We manually interpret reference labels for MFF mapping model training and accuracy assessment. We use the latest pre-flood sub-meter-resolution Esri World Imagery Wayback image, published on June 29, 2023 (Esri, 2023), together with pre-flood 10 m Sentinel-2 RGB imagery to interpret pre-event conditions. Because high-resolution Esri World Imagery is unavailable during the post-flood window, post-flood 10 m Sentinel-2 RGB imagery is compared with pre-flood imagery to identify flood-affected areas and delineate the MFF. The reference labels include standing floodwaters, flood deposits, vegetation stripping and lodging, and structural damage to buildings. To ensure consistency, all delineations are completed by a single interpreter. Based on these data, we select representative severely affected areas as training and validation sites. These sites capture diverse flood evolution patterns, ranging from short-lived mountainous flash floods that receded within one day to

prolonged plain floods lasting nearly one month. Non-flooded sites are also included to assess the ability to suppress false detections in areas unaffected by flooding. In total, the reference labels cover three plain-flood sites, two mountainous-flood sites, and three non-flooded sites.

3. Method

Our framework uses NDWI anomalies as the primary indicator for identifying flood traces, because flood-affected areas can exhibit sustained NDWI anomalies that depart from normal seasonal variability. RGB bands are incorporated as auxiliary constraints to distinguish flood-related NDWI anomalies from non-flood anomalies. The anomalies in the NDWI and RGB bands are then quantified using the CCDC model, in which each observation is compared with the pixel-specific historical baseline and transformed into a residual probability feature. To fully exploit the sustained nature of such anomalies, residual probability features from multiple post-flood dates are then stacked as multi-temporal inputs to a machine-learning classifier for flood trace identification. Finally, to evaluate the effectiveness of the proposed framework, we compare its performance with conventional standing-floodwater-based approaches and conduct ablation experiments from the perspectives of temporal feature selection and spectral band combination (Fig. 2).

3.1. Sustained flood traces and associated NDWI enhancement

Our approach leverages sustained flood traces as the mapping target to reduce MFF underestimation. Beyond the evidence reported in prior studies, we demonstrate and validate the persistence of these traces in 2023 HRB flood event by analyzing multi-source remote-sensing observations. Although apparent flood-induced impacts remain visible in the imagery, most floodwater had already receded by the time Sentinel-1 and cloud-free Sentinel-2 observations became available. As a result, the MFF was difficult to detect using conventional flood-mapping approaches that rely primarily on the presence of standing floodwater (Figs. 3a and 4a). In contrast to the undetectable floodwaters, flood traces remain clearly visible across the affected areas for nearly one month: Severe crop stripping and vegetation damage were observed, while widespread sediment deposits covered bare land, impervious

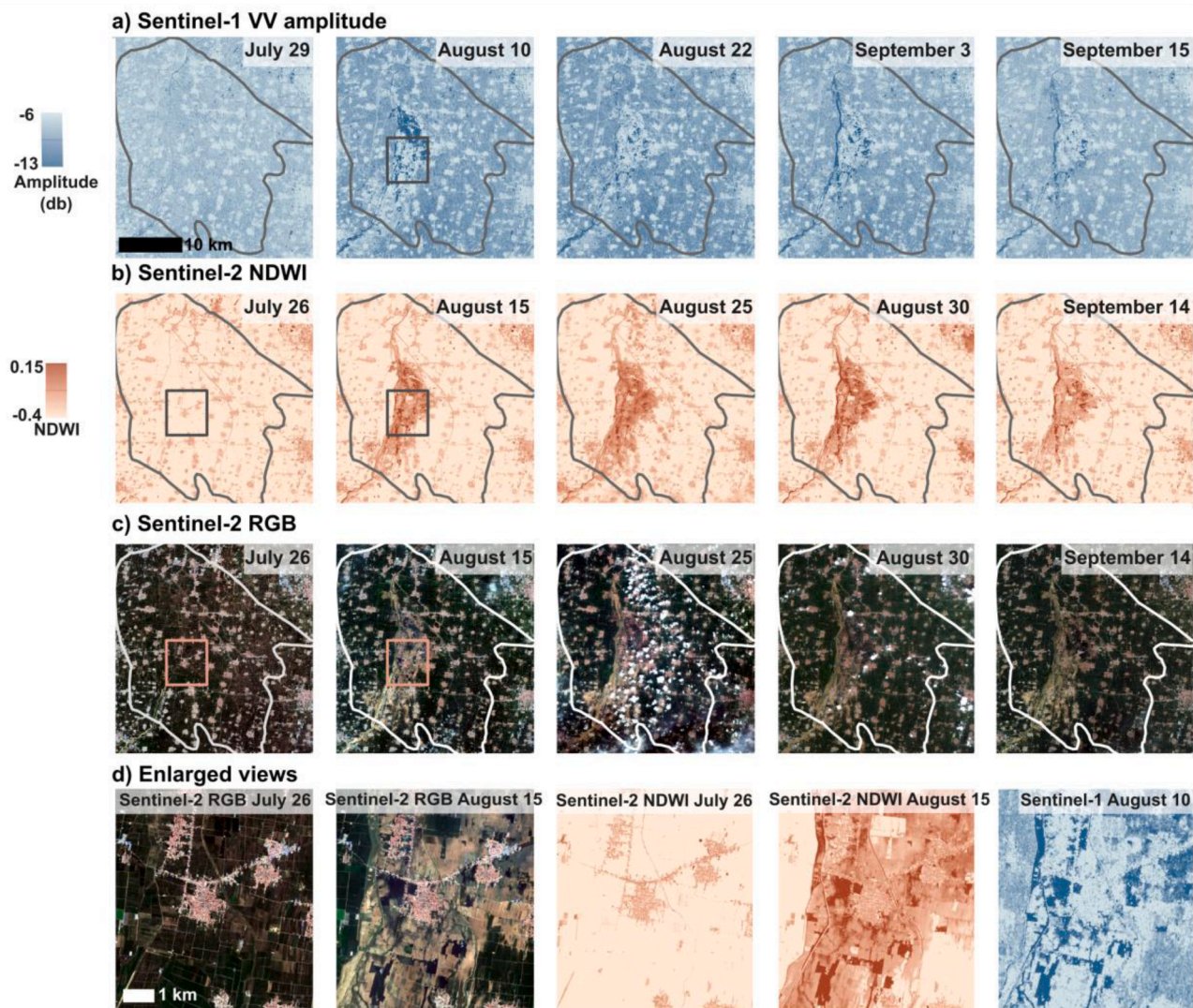


Fig. 3. Multi-source remote sensing imagery illustrating post-flood crop damage and corresponding NDWI anomalies. (a) Sentinel-1 VV-polarized amplitude image. (b) NDWI derived from Sentinel-2 optical imagery. (c) Sentinel-2 true-color composite. (d) Zoomed-in comparisons of pre- and post-flood true-color composites, NDWI, and SAR amplitude imagery in flood affected areas. The geographic locations of the zoomed-in regions are labeled by boxes in their corresponding source images in (a)–(c).

surfaces, damaged vegetation, and cropland (Figs. 3c–d, 4c–d).

Owing to enhanced surface wetness and flood-induced changes in land-surface conditions, the flood traces are accompanied by NDWI elevations. The spatial extent of NDWI enhancement closely matches the MFF inferred from the true-color Sentinel-2 composites, suggesting that NDWI has the potential to serve as an indicator of flood traces (Figs. 3b, d and 4b, d).

3.2. CCDC-based anomaly quantification

Despite the spatial alignment between flood traces and NDWI enhancement, similar NDWI increases can also be caused by non-flood factors, including phenological changes, rainfall-induced wetting, and land-surface changes. For instance, double-cropping harvests may lead to pronounced NDWI increases (Fig. 5b1), while winter vegetation senescence may also increase NDWI (Fig. 5b2–3). Bare land and impervious surfaces may also exhibit relatively high NDWI values (Fig. 5b4–6). In addition, abrupt land-surface changes, such as urban expansion, forest logging, and fire disturbance, may also produce NDWI increases. Therefore, relying only on absolute high NDWI values or simple pre- and post-event NDWI differences may misidentify seasonal processes and

non-flood land-surface changes as flood traces.

Compared with absolute NDWI values or NDWI differences, the degree to which post-flood NDWI deviates from its normal seasonal trajectory provides a more effective indicator for inferring flood traces. This is because flood traces can cause NDWI to exceed its normal seasonal range and remain anomalous across multiple post-flood observation dates. Such anomalies are consistently observed across major land-cover types, including cropland, vegetation, bare land, and impervious surfaces (Fig. 5b). In addition, rainfall events do not produce sustained NDWI anomalies comparable to the post-flood response, thereby preventing rainfall-induced wetting from being misclassified as flooding (Fig. 5d).

To quantify the anomalies, we use the CCDC model to establish a historical normal seasonal trajectory for each pixel. CCDC fits the normal trajectory using temporal trend and seasonal harmonic terms, and identifies potential land-surface change breaks when observations continuously deviate from model predictions. The detected breaks separate the time series into different land-use phases, each described by a distinct set of model coefficients. The CCDC model can be expressed as:

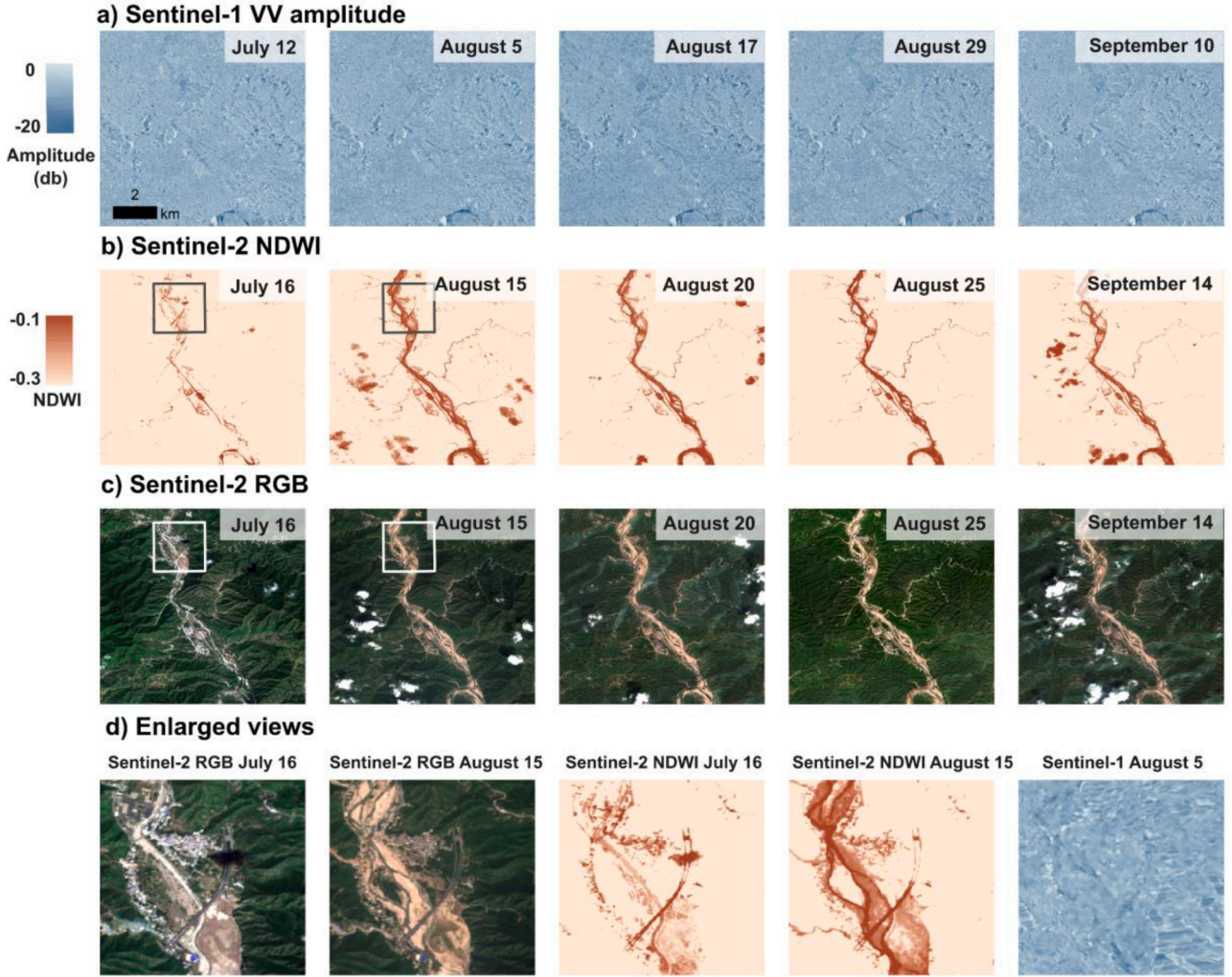


Fig. 4. Multi-source remote sensing imagery illustrating post-flood vegetation stripping, sediment deposition covering bare land and impervious surfaces, and corresponding NDWI anomalies. (a) Sentinel-1 VV-polarized amplitude image. (b) NDWI derived from Sentinel-2 optical imagery. (c) Sentinel-2 true-color composite. (d) Zoomed-in comparisons of pre- and post-flood true-color composites, NDWI, and SAR amplitude imagery in flood affected areas. The geographic locations of the zoomed-in regions are labeled by boxes in their corresponding source images in (a)-(c).

$$\hat{y}(t) = \beta_0 + \beta_1 t + \sum_{k=1}^3 [a_k \cos(2\pi kt) + b_k \sin(2\pi kt)] \quad (1)$$

where $\hat{y}(t)$ is the fitted value at time t , β_0 represents the intercept, and β_1 represents the long-term linear trend. The harmonic terms describe seasonal variations, where $k=1, 2$, and 3 correspond to annual, semi-annual, and tri-annual cycles, respectively. a_k and b_k are the cosine and sine coefficients of the k -th harmonic component, which control the amplitude and phase of the seasonal signal.

CCDC uses two key parameters, *minObservations* and *chiSquareProbability*, to identify land-cover change breakpoints. *ChiSquareProbability* is the chi-square probability threshold used to identify observations that significantly deviate from the fitted time-series model. Observations exceeding this threshold are considered anomalous, and a change breakpoint is detected when such anomalies persist for at least *minObservations* consecutive observations. Based on the sensitivity analysis (see section 5.4.4), *minObservations* is set to 9 and *chiSquareProbability* to 0.85.

The model coefficients from the final land-use segment before the flood event are used to establish a pixel-specific baseline. The deviation of each observation from this baseline is then quantified by the residual:

$$r(t) = y(t) - \hat{y}(t) \quad (2)$$

where $y(t)$ represents an observation, and $\hat{y}(t)$ represents the prediction from the historical normal trajectory.

Because raw residuals are influenced by spatial heterogeneity, they cannot be directly compared across pixels. For instance, the residual distributions differ markedly among land-use types (Fig. 5a). To obtain a unified anomaly metric, we further standardize the raw residuals using the historical residual distribution of each pixel. The historical residuals are assumed to follow a normal distribution:

$$R \sim N(\mu, \sigma^2) \quad (3)$$

where R denotes the residuals, and μ and σ represent the mean and standard deviation of the historical residual distribution, respectively. The residual $r(t)$ can be standardized into residual probability $p(t)$:

$$p(t) = \Phi\left(\frac{r(t) - \mu}{\sigma}\right) \quad (4)$$

where Φ denotes the cumulative distribution function of the standard normal distribution R .

The $p(t)$ represents the percentile position of the post-flood residual within historical residual distribution. A value close to 0.5 indicates that the residual falls near the center of the historical distribution and corresponds to normal variability. A value close to 1 indicates that the re-

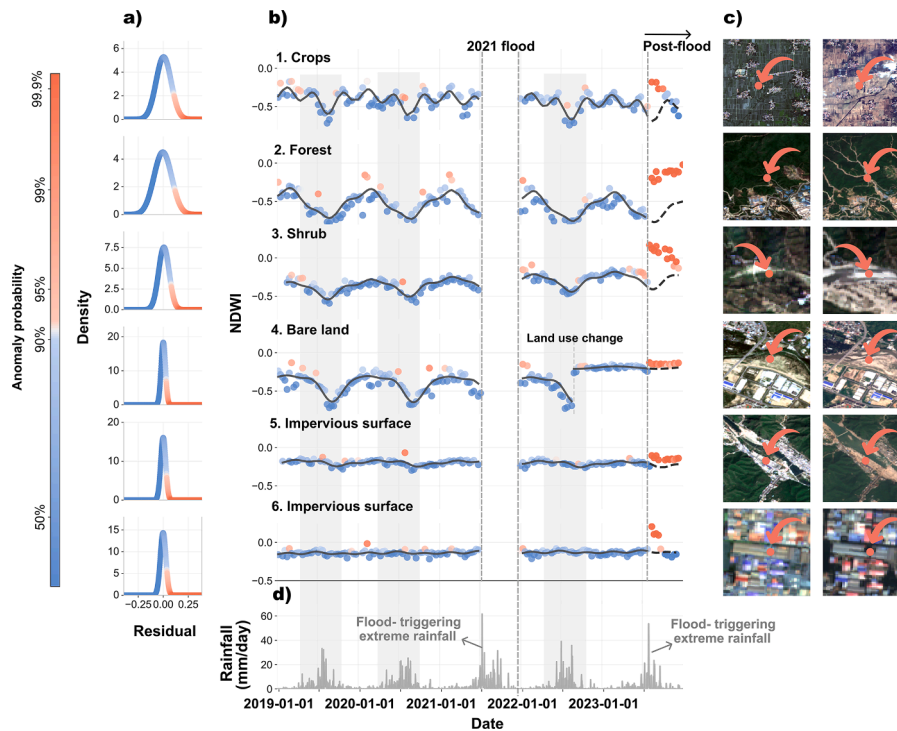


Fig. 5. CCDC-based NDWI anomaly quantification across different land-use types. (a) Historical NDWI residual distributions for different land-use types. Each distribution corresponds to the NDWI time series shown in the same row of panel (b), and the pre- and post-flood Sentinel-2 RGB images in panel (c). (b) CCDC-fitted historical trajectories and post-flood NDWI observations across different land-use types. The color of each observation represents its residual probability, as indicated in panel (a). Solid black lines indicate CCDC-fitted historical trajectories, while dashed black lines indicate CCDC-based post-flood predictions. (c) Pre- and post-flood Sentinel-2 RGB images. (d) Daily rainfall time series, with flood-triggering extreme rainfall events marked. The daily rainfall was derived from the CHIRPS daily precipitation dataset (Funk et al., 2015) and calculated as the spatial mean precipitation over the study area for each day. The rainfall values are reported in millimeters per day (mm/day).

sidual falls in the upper tail of the historical distribution, representing a strong positive anomaly. In contrast, a value close to 0 indicates that the residual falls in the lower tail of the historical distribution, representing a strong negative anomaly.

The probability maps residuals from different pixels into a unified probability space, making anomaly intensity comparable across land-surface types. Therefore, CCDC-based residual probability features can effectively reduce the interference of normal seasonal fluctuations and inter-pixel heterogeneity, providing a unified anomaly features for subsequent machine-learning classification.

3.3. MFF mapping

3.3.1. Feature dataset integrating NDWI and RGB anomalies

Although CCDC-based anomaly quantification of NDWI can reduce the effects of phenological changes and rainfall-induced wetting, some non-flood land-surface changes, such as urban expansion, forest logging, land clearing, and pest damage, may still cause NDWI to deviate from its normal seasonal trajectory. For instance, abrupt land-use conversion from vegetation to bare land can also generate NDWI anomalies (Fig. 5b4). Therefore, relying only on NDWI anomalies to map flood traces may misclassify certain surface changes as flood-related responses. To reduce confusion caused by such changes, we further introduce the R, G, and B anomalies as auxiliary constraints. RGB bands provide complementary information on surface color, brightness, and land-cover conditions in the visible spectrum, helping determine whether an NDWI anomaly is accompanied by visible-spectrum responses consistent with flood traces. Accordingly, CCDC-based anomalies are also calculated separately for the red, green, and blue bands. The NDWI and RGB anomaly features are then jointly used to construct the feature dataset for machine-learning-based flood trace classification.

To better exploit the persistence of flood traces, which can remain detectable for over a month after the flood event in HRB, we extract anomaly features from four post-flood time steps for NDWI, red, green, and blue bands. The four variables across four time-steps are stacked to construct a 16-band feature dataset for subsequent machine-learning classification.

3.3.2. Machine learning algorithm training

Given that flood traces may exhibit different spectral responses under different geomorphic conditions, we include both plain-flood and mountainous-flood sites when training and validating the machine-learning model. Non-flooded sites are also included to constrain false detections. We adopt a site-level split between training and testing datasets. Specifically, one mountainous-flood site, one plain-flood site, and one non-flooded site are reserved as independent testing sites, while the remaining sites in each category are used for model training. Compared with randomly splitting samples from all sites into training and testing sets, this site-level split ensures spatial independence between training and testing data, reduces the influence of spatial autocorrelation and potential information leakage, and provides a more robust assessment of model performance in unseen regions.

To determine the machine-learning classification scheme, we evaluate different combinations of model type, hyperparameters, input feature sets, and positive-to-negative training sample ratios within a cross-validation framework. Three models are compared: Random Forest (RF), XGBoost, and Deep Neural Network (DNN). Three input feature sets are evaluated: combined RGB and NDWI residual probabilities, NDWI residual probability only, and RGB residual probabilities only. Positive-to-negative training sample ratios ranging from 1:1 to 1:5 are also examined.

The cross-validation strategy accounts for spatial autocorrelation

Table 1

Cross-validation results of three machine learning models.

Performance (%)	Precision	Recall	F1	IoU	Balanced accuracy
XGBoost	79.44 ± 4.10	88.59 ± 2.58	83.71 ± 2.88	72.09 ± 4.30	92.65 ± 1.42
RF	83.02 ± 4.36	84.41 ± 3.71	83.66 ± 3.45	72.04 ± 5.08	90.96 ± 1.98
DNN	80.12 ± 3.47	82.37 ± 5.92	81.11 ± 3.57	68.36 ± 5.15	89.73 ± 2.90

among neighboring remote-sensing pixels. Random pixel-based partitioning may place spatially adjacent and highly similar samples in both the training and validation sets, resulting in information leakage and overestimation of model performance (Roberts et al., 2017). To mitigate this issue, we partition the training regions into spatial blocks of 100×100 pixels ($1 \text{ km} \times 1 \text{ km}$ at 10-m Sentinel-2 resolution), and then apply repeated five-fold spatial cross-validation with two repeats, yielding 10 validation runs for each candidate configuration. To ensure consistent comparisons, all candidate configurations use the same spatial blocks and cross-validation partitions. Compared with a single 10-fold cross-validation, the repeated five-fold strategy retains a relatively large validation sample size in each fold while enabling further assessment of model sensitivity to different spatial training-validation partitions.

Model performance is evaluated using multiple metrics, including Precision, Recall, F1 score, Intersection over Union (IoU), and Balanced Accuracy. For each candidate configuration, these metrics are calculated separately for the 10 spatial cross-validation runs and summarized as the mean $\pm$ standard deviation (mean $\pm$ SD). The mean represents the overall performance across different spatial validation subsets, whereas the standard deviation reflects the stability of model performance across different spatial data partitions. We convert the machine-learning mapping results into vector polygons and further apply a 50 m Height Above Nearest Drainage (HAND) threshold following Nobre et al. (2011) to reduce false detections caused by mountainous terrain shadows and cloud contamination.

3.4. Validation, benchmarking, and ablation

3.4.1. Benchmark against floodwater-based method

To demonstrate the advantages of our framework, we compare its results with standing-floodwater-based mapping methods using different combinations of input data. To ensure consistency across these baselines, we adopt the udl4fl framework proposed by Portalés-Julìà et al. (2025). The framework provides deep-learning flood segmentation models for three input modalities: Sentinel-1 SAR, Sentinel-2 multispectral imagery, and their multimodal combination. By leveraging diverse training samples from multiple benchmark flood datasets, udl4fl is designed to improve the robustness and generalizability of flood mapping across different regions and sensor modalities. The MFF for each data combination is generated by spatially unioning the corresponding multi-temporal flood-mapping results.

- (1) Multi-temporal Sentinel-1 images: We use the corresponding single-modality udl4fl model trained exclusively on Sentinel-1 observations. Since simple differencing of pre- and post-flood water extents can be affected by seasonal variations in surface-water extent (Lin et al., 2025; Martinis et al., 2022), we further use the Reference Water Mask provided by the Global Flood Monitoring (GFM) service to remove permanent and seasonal water bodies from the mapped flood extent. We further filter the mapped flood extents using the GFM Exclusion Mask to improve the reliability of the standing-floodwater-based results in areas affected by factors such as dense vegetation, urban surfaces, and unfavorable topographic conditions (Wagner et al., 2026).
- (2) Multi-temporal Sentinel-2 images: Similarly, the MFF derived from Sentinel-2 imagery is generated using the single-modality udl4fl model trained exclusively on Sentinel-2 observations. The seasonal water bodies are removed using the OPERA Dynamic Surface Water Extent product (OPERA DSWx-HLS; OPERA, 2023).
- (3) Multi-source SAR integration: Since underestimation of MFF from single-source imagery has been observed, commercial SAR data are often incorporated to mitigate this issue. Therefore, we further integrate the flood extents derived from Gaofen-3 and Lutan-1 SAR imagery with Sentinel-1 results. For dual-polarized

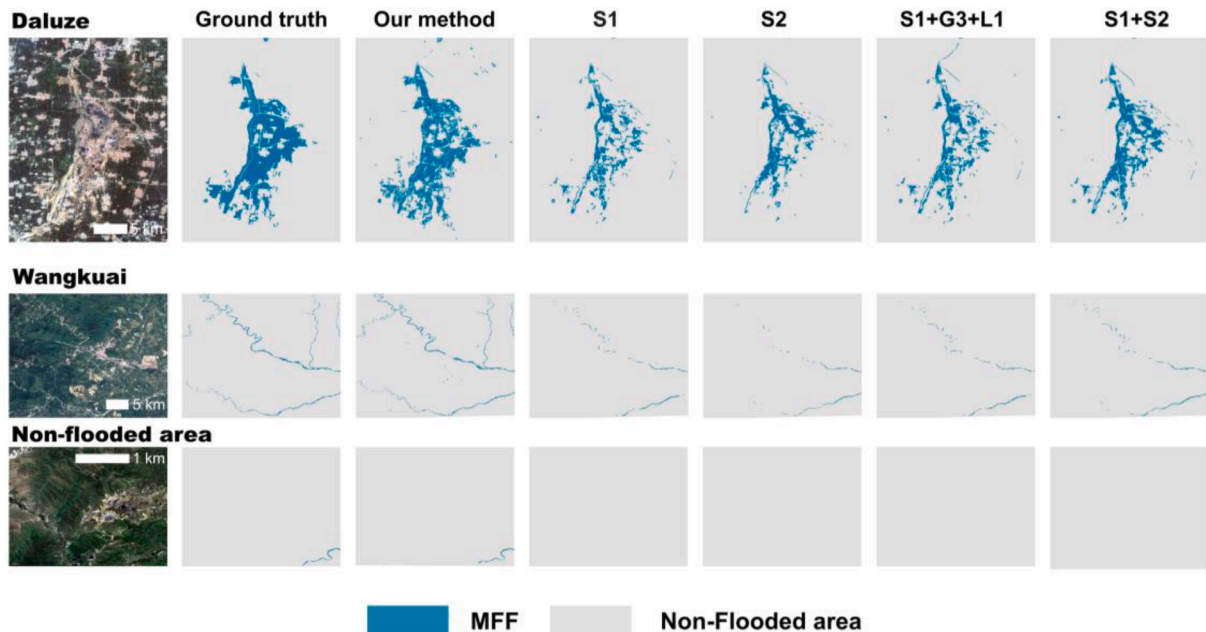


Fig. 6. Visual comparison against reference labels for our method and four commonly used methods. S1, S2, G3, and L1 denote Sentinel-1, Sentinel-2, Gaofen-3, and Lutan-1, respectively.

Table 2

Quantitative comparison against reference labels for our method and four commonly used flood-mapping methods.

Method	Performance (%)	Daluze	Wangkui	Non-flooded area	Average
Ours	Precision	84.65	76.05	78.45	79.72
	Recall	81.18	88.25	81.12	83.52
	F1	82.88	81.70	79.76	81.45
	IoU	70.76	69.06	66.33	68.72
	Balanced accuracy	89.63	93.92	90.54	91.36
Sentinel-1	Precision	93.22	86.02	0.00	59.75
	Recall	40.05	24.84	0.00	21.63
	F1	56.03	38.55	0.00	31.52
	IoU	38.92	23.87	0.00	20.93
	Balanced accuracy	69.83	62.39	50.00	60.74
Sentinel-2	Precision	93.45	84.49	0.00	59.31
	Recall	38.26	18.55	0.00	18.94
	F1	54.29	30.42	0.00	28.24
	IoU	37.26	17.94	0.00	18.40
	Balanced accuracy	68.95	59.25	50.00	59.40
Sentinel-1 +Gaofen-3 +Lutan-1	Precision	92.16	86.02	0.00	59.39
	Recall	43.22	24.84	0.00	22.69
	F1	58.84	38.55	0.00	32.46
	IoU	41.69	23.87	0.00	21.85
	Balanced accuracy	71.37	62.39	50.00	61.25
Sentinel-1 +Sentinel-2	Precision	92.85	84.82	0.00	59.22
	Recall	50.71	29.38	0.00	26.69
	F1	65.59	43.64	0.00	36.41
	IoU	48.80	27.91	0.00	25.57
	Balanced accuracy	75.10	64.65	50.00	63.25

Gaofen-3, we synthesize a false-color RGB composite and feed it into DeepLabV3+. To harmonize cross-sensor radiometry and enhance generalizability, we apply histogram matching to the composite prior to inference (Lin et al., 2025). For single-polarized Lutan-1, we employ Otsu thresholding to extract the flood extent.

- (4) Fusion of Sentinel-1 and Sentinel-2: Sentinel-1 SAR and Sentinel-2 optical imagery provide complementary information and are therefore widely used in combination for flood mapping (e.g., Ahmadi et al., 2025, 2026). Because multimodal fusion requires close temporal correspondence between SAR and optical observations, we pair Sentinel-1 and Sentinel-2 images acquired on the same day or on consecutive days and use them as inputs to UDL4FL. When temporally matched image pairs are unavailable, Sentinel-1 and Sentinel-2 observations are processed separately using the corresponding single-sensor methods described above. Finally, the multimodal fusion results and the independently derived Sentinel-1 and Sentinel-2 flood maps are merged through a union operation to obtain the final flood extent. To ensure a consistent comparison, the results generated by our framework are also filtered using OPERA DSWx-HLS.

3.4.2. Ablation experiments

To analyze the contribution of different features and demonstrate the necessity of using multi-temporal and multi-band data, we conduct both spectral-feature ablation and temporal-feature ablation experiments. To ensure fair comparison, all ablation experiments use the same training-validation split and machine-learning model.

The spectral-feature ablation experiment evaluates the different roles of NDWI and RGB anomaly features in MFF identification. The experiment compares three configurations: RGB-only residual

probability features, NDWI-only residual probability features, and the full NDWI-RGB residual probability feature set adopted in our framework.

In addition to the spectral dimension, we also conduct a temporal-feature ablation experiment to evaluate the contribution of the proposed multi-temporal CCDC-based anomaly features. This experiment compares three temporal feature configurations: The first configuration uses the absolute RGB and NDWI values from a single post-flood time step, representing a conventional single-date post-flood feature setting. The second configuration uses the pre- and post-flood differences of absolute RGB and NDWI values, representing a widely-used before-after change feature setting. The third configuration is the proposed multi-temporal anomaly feature setting, which stacks CCDC-based residual probability features of NDWI and RGB from multiple post-flood time steps.

3.5. Uncertainty estimation

3.5.1. Manual reference label interpretation

Manual interpretation can involve uncertainty because flood traces often exhibit gradual transitions and difficulties in confidently identifying some flood-affected areas. To quantify this uncertainty, one representative mountainous site and one representative plain site are independently delineated twice by the interpreter. To minimize memory effects and improve the independence of the repeated interpretations, the two delineations are separated by four months. Area-level agreement between the two delineations is quantified using the F1 score and intersection over union (IoU). To assess transition-zone uncertainty, corresponding boundary segments from the two delineations are paired, and points are sampled at 10-m intervals along one segment to calculate their distances to the nearest points on the paired segment.

3.5.2. Flood-trace persistence

Since our method assumes that flood traces persist until cloud-free optical imagery becomes available, rapidly fading traces may introduce uncertainty. To assess this uncertainty, we use independent high-resolution data to evaluate flood-trace persistence across different land-use types and its effects on reference-label collection and MFF mapping. For impervious surfaces, buildings, and shrub, we use the MFF ground truth for Hantai Village, Changping District, Beijing, provided by Ma et al. (2025), derived from high-resolution UAV imagery and detailed field surveys. The field surveys capture building-interior damage that is not visible in satellite imagery, providing a more reliable estimate of MFF in built-up areas. For forests and farmland, we delineate MFF using high-resolution UAV imagery of Zhuozhou City, Hebei Province, provided by Wu et al. (2024). Because inundation beneath buildings cannot be directly observed from UAV imagery, we do not derive MFF-ground-truth over impervious surfaces and buildings in Zhuozhou. We compare the high-resolution MFF ground truth with Sentinel-2-derived reference labels to assess the effect of flood-trace persistence on label collection. We do not analyze bare land because we could not identify reliable independent high-resolution MFF data for this land-use type.

We further quantify the persistence duration of flood traces across different land-use types. For each land-use type, we collect time series of NDWI residual probabilities from 13 August to 31 December, 2023. We adopt a threshold of 0.85 for the NDWI residual probability, consistent with the CCDC breakpoint criterion used in our model. Dates with residual probabilities above 0.85 are considered anomalous and detectable by the model. Flood-trace persistence duration is then defined as the time required for the NDWI residual probability to fall below this threshold.

3.5.3. Cloud-induced observation gap

Our method relies on post-flood cloud-free optical observations. Observation gaps caused by clouds or other noise represent another

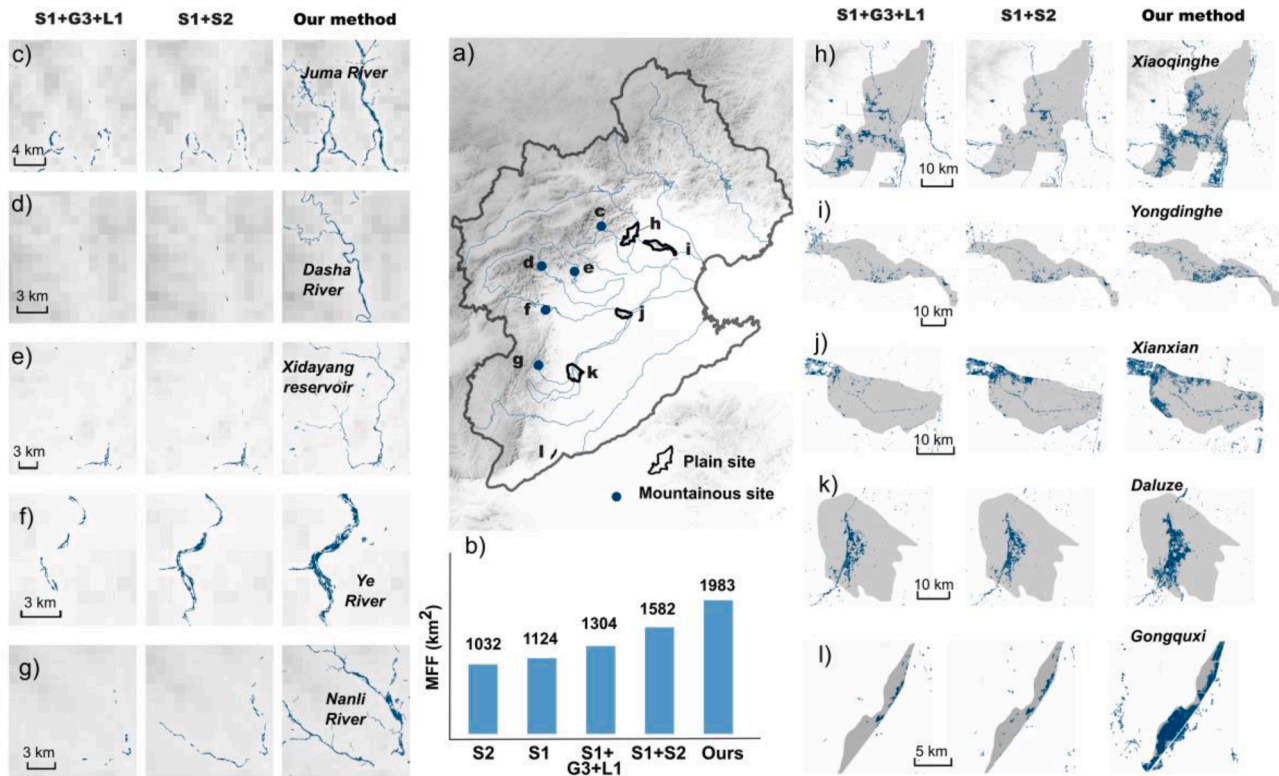


Fig. 7. MFF of 2023 Hai River Basin flood. (a) Locations of representative sites in both plain-area and mountainous areas demonstrating the mapping results of different methods. (b) The total mapped MFF area for each method. (c-g) Detailed comparison between our method and classic methods in mountainous regions. (h-i) Detailed comparison in activated FDAs across the eastern plains.

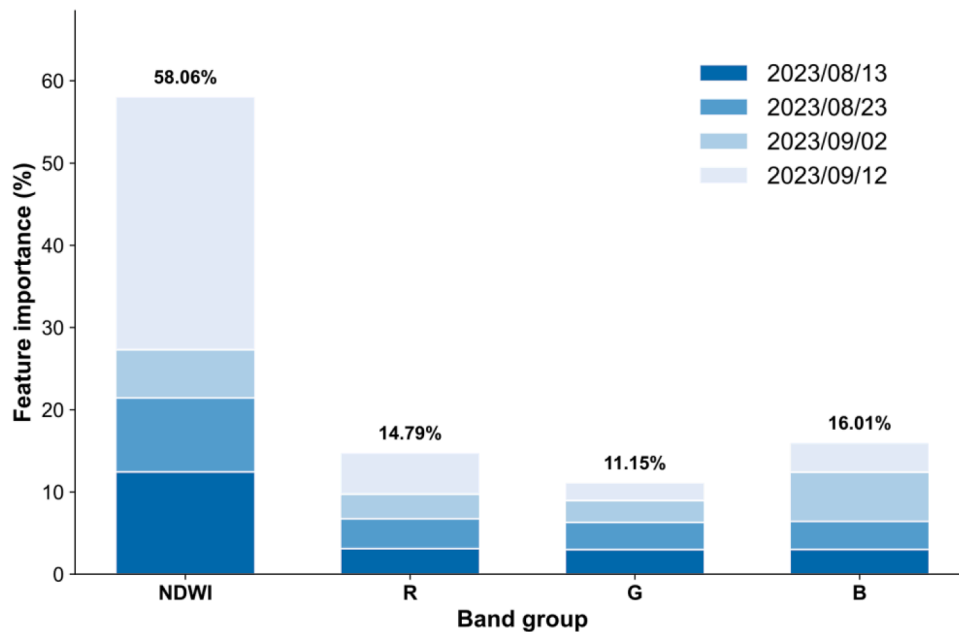


Fig. 8. Feature importance of NDWI and RGB anomaly features.

source of uncertainty. To quantify the effect of fewer clear observations on MFF mapping, we conduct a controlled missing-observation stress test. Since the machine-learning model does not explicitly use acquisition dates as temporal information, missing observations from different dates may lead to different mapping accuracies even when the number of missing observations is the same. Therefore, rather than considering only the number of missing observations, we evaluate all possible

combinations of missing observations among the four post-flood dates, resulting in 14 missing-observation scenarios: four with one missing date, six with two missing dates, and four with three missing dates, while the complete four-date case serves as the baseline. For each scenario, we artificially construct missing observations in the testing dataset by setting the RGB and NDWI residual-probability features of the corresponding dates to missing values. The modified testing datasets are

Table 3
Ablation results for different feature configurations.

Method	Performance (%)	Daluze	Wangkuai	Non-flooded area	Average
Ours	Precision	84.65	76.05	78.45	79.72
	Recall	81.18	88.25	81.12	83.52
	F1	82.88	81.70	79.76	81.45
	IoU	70.76	69.06	66.33	68.72
	Balanced accuracy	89.63	93.92	90.54	91.36
NDWI-only residual probability	Precision	85.96	75.82	41.44	67.74
	Recall	80.41	87.16	90.30	85.96
	F1	83.09	81.10	56.81	73.67
	IoU	71.08	68.20	39.67	59.65
	Balanced accuracy	89.35	93.38	95.04	92.59
RGB-only residual probability	Precision	84.51	43.34	75.53	67.79
	Recall	47.92	62.51	40.26	50.23
	F1	61.16	51.19	52.52	54.96
	IoU	44.05	34.40	35.62	38.02
	Balanced accuracy	73.39	80.65	70.12	74.72
Post-flood feature	Precision	40.17	25.04	4.70	23.31
	Recall	72.08	89.80	86.96	82.95
	F1	51.59	39.17	8.92	33.23
	IoU	34.76	24.35	4.67	21.26
	Balanced accuracy	79.03	92.91	91.85	87.93
Before-after change feature	Precision	90.45	71.89	38.24	66.86
	Recall	60.66	85.72	93.04	79.80
	F1	72.62	78.20	54.20	68.34
	IoU	57.01	64.20	37.18	52.80
	Balanced accuracy	79.91	92.61	96.38	89.63

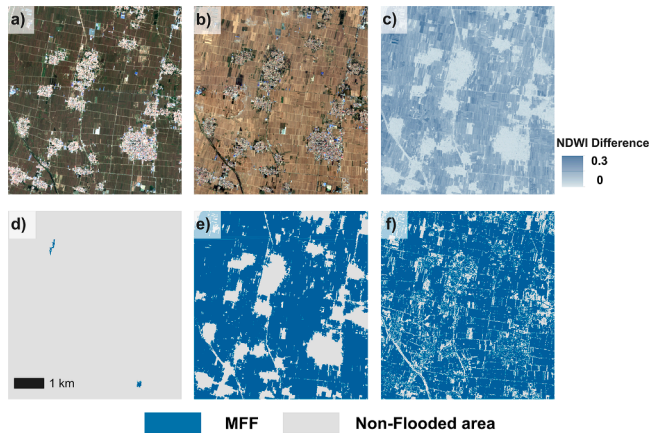


Fig. 9. Comparison of MFF results for a non-flood harvest-season case in the Daluze FDA. (a) Sentinel-2 RGB image before harvesting on June 5, 2023. (b) Sentinel-2 RGB image after harvesting on June 25, 2023. (c) NDWI difference between the post-harvest and pre-harvest images. (d) MFF identification result from the historical-baseline-based anomaly feature setting. (e) Result from the before-after change feature setting. (f) Result from the single-date post-flood feature setting.

then input into the machine-learning model, and their performance is compared with the complete-observation baseline.

4. Results

4.1. Performance in validation sites

The cross-validation results show that XGBoost with a positive-to-

negative training sample ratio of 1:3 achieves the best performance, with an F1 score of $83.71\% \pm 2.88\%$, outperforming Random Forest ($83.66\% \pm 3.45\%$) and DNN ($81.11\% \pm 3.57\%$) (Table 1). The optimal hyperparameter configuration for each model is provided in Table S1. Therefore, XGBoost is adopted for subsequent method comparisons and catchment-scale applications.

Comparison with conventional standing-floodwater-based methods shows that our method maintains consistently strong performance, outperforming commonly used flood-mapping approaches. The mapped MFF closely match the manually interpreted reference labels (Fig. 6), with average F1 scores of 81.45%. Moreover, the performance remains stable across plain-flood, mountainous-flood, and non-flooded validation sites, indicating that the proposed framework can capture MFF under different geomorphic conditions and suppress false detections in areas unaffected by flooding. In contrast, the four commonly used methods substantially underestimate MFF across all validation sites, with F1 scores only ranging from 28.24% to 36.41%. This underestimation is primarily due to the limited temporal availability of remote-sensing observations, as much of the floodwater had already receded by the time the images were acquired (Table 2).

4.2. Validation at catchment scale

At the catchment scale, our method extracts a larger MFF than the commonly used methods, retrieving 1,983 km² of MFF. By contrast, flood mapping based solely on Sentinel-1 and Sentinel-2 identifies only 1124 km² and 1032 km², respectively, 43.3% and 48.0% smaller than our estimate. Incorporating commercial SAR imagery only marginally reduces this gap because of its limited spatial and temporal coverage, increasing the mapped MFF to 1304 km², which remains 34.2% smaller than our estimate. The Sentinel-1 and Sentinel-2 fusion yields an MFF of 1,582 km², which is approximately 20.2% smaller than our estimate. Fusing Sentinel-1 and Sentinel-2 alleviates the underestimation of MFF by leveraging their complementary sensing capabilities, which increase the availability of valid observations and reduce flood omissions caused by cloud contamination and sensor-specific observation limitations (Fig. 7a-b).

The advantages of our approach are most pronounced in the western mountainous sector. All methods relying on standing-floodwaters fail to capture the expansion of river channels during the 2023 HRB flood events, whereas our approach detects these changes that are subtle in space and brief in duration (Fig. 7c-g). In the eastern plains, the longer-lasting waterlogging in FDAs reduces the degree of underestimation. Nonetheless, substantial discrepancies persist (Fig. 7h-i), particularly in Gongquxi FDA (Fig. 7i).

5. Discussion

5.1. The central role of NDWI anomalies

Feature importance results indicate that NDWI anomalies serve as the primary indicator for identifying MFF. The total importance of NDWI anomaly features reaches 58.06%, substantially exceeding those of the R, G, and B anomaly features, which account for 14.79%, 11.15%, and 16.01%, respectively (Fig. 8). This dominance is further supported by the spectral-feature ablation experiment. The NDWI-only configuration achieves an average F1-score of 73.67%, approaching the full NDWI-RGB feature combination (81.45%) and markedly higher than the RGB-only configuration (54.96%) (Table 3). These results demonstrate that NDWI anomalies provide the dominant discriminative information for MFF identification.

Nevertheless, NDWI anomalies alone are not sufficient to fully suppress false detections. In the non-flooded area, the precision of the NDWI-only configuration is 41.44%, lower than the 78.45% achieved by the full feature combination. Incorporating RGB anomalies improves the average F1-score across all validation sites from 73.67% to 81.45%.

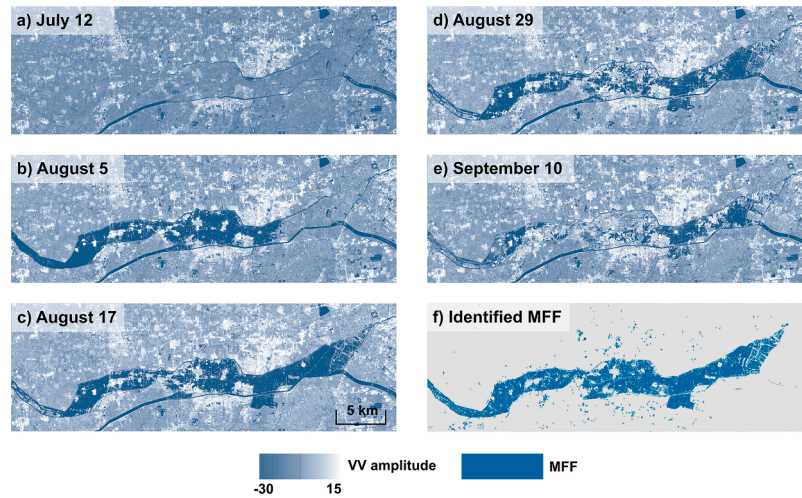


Fig. 10. Standing floodwater and identified MFF in the Dongdian FDA. (a–e) Multi-temporal SAR VV amplitude images, where low VV amplitude values indicate standing floodwater. (f) MFF identified by the proposed method.

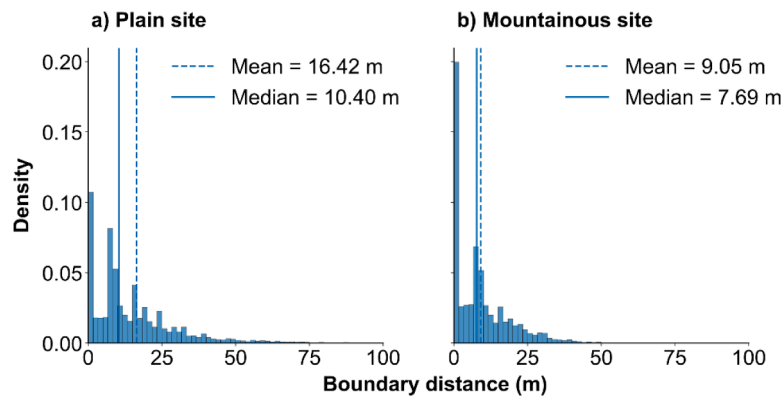


Fig. 11. Distribution of transition-zone boundary distances at the plain and mountainous sites.

Table 4
Quantitative uncertainty assessment of manual reference-label interpretation.

Site	Precision	Recall	F1	IoU	Average transition zone uncertainty (m)
Plain site	95.61	92.63	94.10	88.85	16.42
Mountainous site	90.60	92.20	91.39	84.15	9.05
Average	93.11	92.42	92.75	86.50	12.74

Therefore, although RGB anomalies are not the primary information source, they provide effective auxiliary constraints for distinguishing flood-related anomalies from non-flood disturbances, thereby improving the reliability of MFF mapping (Table 3).

5.2. The necessity of incorporating historical baselines

The temporal-feature ablation experiment highlights the importance of incorporating historical baseline. The proposed historical-baseline-based anomaly features achieve an average F1-score of 81.45%, outperforming the single-date post-flood feature setting and the before-after change feature setting by 48.22 and 13.11 percentage points, respectively. This advantage is particularly pronounced in non-flooded areas: Our method achieves an F1-score of 79.76%, whereas the single-date post-flood and before-after change settings reach only 8.92% and 54.20%, respectively (Table 3).

We further demonstrate this necessity by applying the three feature

settings to a non-flood harvest-season case in the Daluze FDA in June 2023. Harvesting substantially alters vegetation cover and exposes bare soil, which can generate large before-after spectral differences, especially pronounced NDWI enhancement (Fig. 9a–c). As a result, both the single-date post-flood feature setting and the before-after change feature setting incorrectly identify large areas of harvested cropland as MFF (Fig. 9e–f). In contrast, the historical-baseline-based anomaly features distinguish MFF from normal seasonal agricultural changes and produce almost no false positives (Fig. 9d). These results indicate that historical baselines are essential for distinguishing MFF from normal seasonal or land-use-related surface changes.

5.3. Mapping both standing floodwater and flood traces

Although the proposed method primarily leverages sustained flood traces to identify MFF, it can also detect standing floodwater that remains present when satellite observations become available. For example, in the Dongdian FDA, substantial standing floodwater persists across the four post-flood feature time steps from August 13 to September 12 (Fig. 10a–e), and these areas are correctly identified as MFF by the proposed method (Fig. 10f). This is because standing floodwater also produces strong NDWI anomalies and visible-spectrum responses. In addition, within the regions used to generate training samples, pixels covered by standing floodwater are also included in the MFF labels.

This characteristic prevents the method from missing prolonged waterlogging areas due to an exclusive focus on flood traces. More

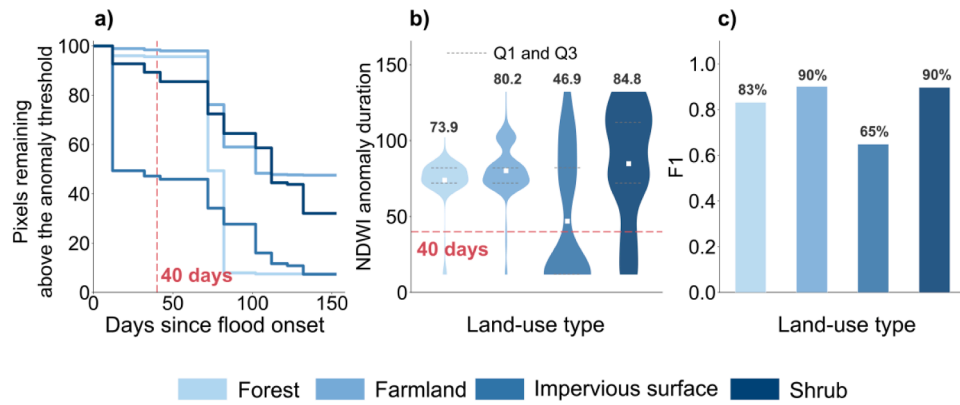


Fig. 12. Flood-trace persistence quantification across different land-use types. (a) Proportion of pixels remaining above the anomaly threshold for different land-use types. (b) Distribution of anomaly persistence durations for different land-use types. (c) F1 scores between the Sentinel-2-derived reference labels and the high-resolution MFF ground truth.

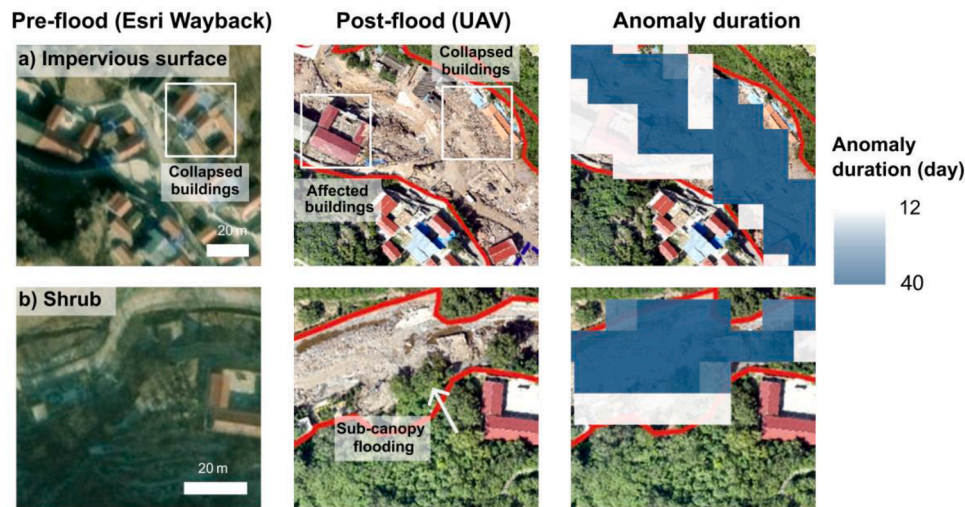


Fig. 13. Anomaly duration over shrub and impervious surfaces illustrated using pre-flood Esri Wayback imagery, post-flood UAV imagery, and pixel-level anomaly-duration maps.

importantly, by capturing both remaining standing floodwater and sustained post-flood traces, the proposed method maintains a conservative baseline comparable to conventional standing-floodwater-based mapping even when flood traces are weak or short-lived. In most cases, however, flood traces remain detectable after water recession, allowing the method to further expand the recoverable MFF beyond this baseline.

5.4. Uncertainty estimation

5.4.1. Manual reference label interpretation

Uncertainty in reference-label interpretation remains relatively limited. The mean F1 score and IoU between the two interpretations are 92.75% and 86.50%, respectively. The mean boundary difference within the transition zones is 12.74 m, corresponding to ~ 1.3 Sentinel-2 pixels. Most transition-zone differences at both the plain and mountainous sites are below 25 m, equivalent to ~ 2.5 Sentinel-2 pixels (Fig. 11 and Table 4).

5.4.2. Flood-trace persistence

The MFF ground truth derived from high-resolution data demonstrates that the assumption that flood traces can be persistent holds for most areas. The majority of cropland, forest, and shrub pixels exhibit anomalous NDWI signals for more than 40 days, and even ~ 150 days

after the flood, a substantial proportion of cropland and shrubland pixels remain anomalous (Fig. 12a). The mean persistence durations of NDWI anomalies are 73.9, 80.2, and 84.8 days for forests, croplands, and shrublands, respectively (Fig. 12b). Such persistent anomaly signals allow post-flood optical observations to capture flood traces over extended periods, enabling the model to identify most MFF areas effectively. They also support the reliability of Sentinel-2-derived reference labels for these land-use types: compared with the high-resolution MFF ground truth, the reference labels achieve F1 scores of 83%, 90%, and 90% for forests, croplands, and shrublands, respectively (Fig. 12c).

Despite the good agreement for croplands, forests, and shrublands, the main uncertainty associated with flood-trace persistence occurs over impervious surfaces. A proportion of impervious-surface pixels lose their anomalous NDWI signals within 40 days after the flood (Fig. 12a-b), and the agreement between the high-resolution MFF ground truth and the Sentinel-2-derived reference labels is relatively low, with an F1 score of only 65% (Fig. 12c). These rapidly fading anomalies mainly occur over buildings that were inundated internally but remained structurally intact. Because their roofs show little or no visible flood traces, they cannot be reliably identified from overhead remote-sensing imagery. In contrast, buildings damaged or destroyed by flooding exhibit clear sediment and damage signatures, producing strong and persistent anomalies that can be detected (Fig. 13a). A small fraction of flood traces

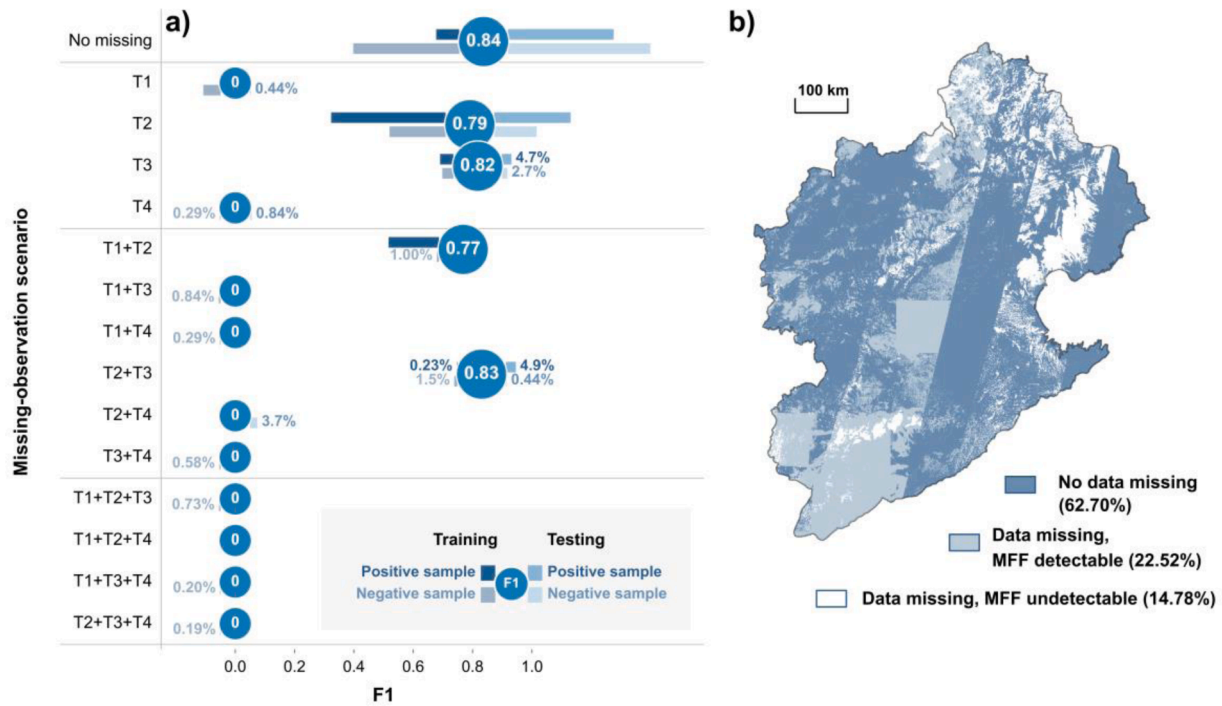


Fig. 14. Missing-observation stress test. (a) Model performance under different missing-observation scenarios. (b) Spatial distribution of different missing-observation scenarios across the HRB.

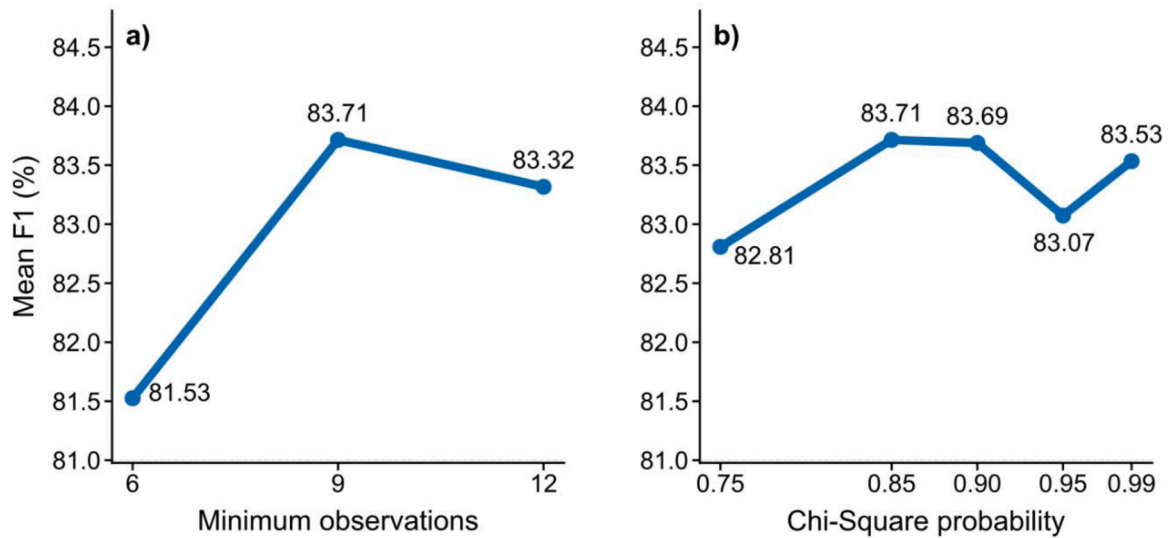


Fig. 15. Sensitivity of model performance to CCDC parameters. (a) Sensitivity to the minimum number of observations. (b) Sensitivity to the chi-square probability.

beneath shrub vegetation are similarly missed (Figs. 12 and 13a). Comparable omissions are not evident in the cropland and forest assessments because the high-resolution UAV data used to derive ground truth are themselves unable to observe flood traces obscured by dense vegetation (Fig. 13b).

5.4.3. Cloud-induced observation gap

The stress-test results demonstrate that the model exhibits a certain degree of robustness to partial data loss caused by clouds and other sources of noise. This robustness is primarily attributable to the fact that the training data already contain several missing-observation scenarios. Under these scenarios, the model maintains F1 scores ranging from 77% to 83%, which are close to the performance obtained with complete four-date observations. In contrast, for missing-observation scenarios

not represented in the training data, the model may lose its ability to identify MFF (Fig. 14a).

However, these unseen missing-observation scenarios do not substantially compromise the large-scale applicability of the model in HRB because we have selected representative training and testing datasets. Across the HRB, 62.70% of the area has complete observations for all four dates, while another 22.52% corresponds to missing-observation scenarios already represented in the training dataset and can be effectively identified. Only 14.78% of the area corresponds to missing-observation scenarios not represented in the training data, and these areas are mainly distributed in the northeastern part of the HRB, outside the core flood-affected region. In addition, the testing dataset includes three missing-observation scenarios, ensuring the representativeness and reliability of the evaluation (Fig. 14a and b). The stress test indicates

that, when constructing training datasets for large-scale flood mapping, representativeness should be considered not only in terms of geographic characteristics but also with respect to possible missing-observation scenarios. Incorporating these scenarios can improve model robustness under incomplete observation conditions.

5.4.4. Sensitivity to model parameters

The model performance remains generally stable across different CCDC break detection parameters. The mean F1 score varies only moderately with changes in both the minimum number of observations and the chi-square probability. The best performance is achieved with a minimum observation number of 9 and a chi-square probability of 0.85, yielding a mean F1 score of 83.71% in the cross-validation analysis. Therefore, these two values are adopted as the final CCDC parameters (Fig. 15). The framework's performance also remains stable across different hyperparameter settings. When using the same input feature set and positive-to-negative training sample ratio as the optimal model, the F1 score of XGBoost ranges from $81.21\% \pm 2.98\%$ to $83.71\% \pm 2.88\%$ across different hyperparameter configurations.

5.5. Limitation and future work

Despite demonstrating advantages over existing flood-mapping methods during the 2023 HRB flood event, our approach still has several limitations. First, remote-sensing observations cannot directly capture flood traces inside buildings or beneath dense vegetation, which may lead to local underestimation of MFF. The limited spatial resolution of satellite imagery may hinder reliable MFF delineation in transition zones, where flooded and non-flooded surfaces are spatially mixed. Future work could integrate hydrodynamic models or terrain-based gap-filling approaches to reduce such underestimation (Li et al., 2025; Bates et al., 2021). Second, although the model demonstrates a certain degree of robustness to missing observations, it fails to detect flood traces under missing-data scenarios that are not represented in the training data. This limitation is partly attributable to the lack of temporal information in the current machine-learning model. Future work could incorporate temporal models and further fuse optical, SAR, and other complementary data to reduce the effects of cloud contamination and improve discrimination between flood-related and non-flood disturbances (Ahmadi et al., 2025). Third, CCDC requires long time series to construct historical baselines, resulting in relatively high computational demands for large-scale applications. Future improvements in time-series optimization and parallel computing could enhance the scalability of the framework. Finally, the current study evaluates the method using a single representative flood event. Future work should extend the framework to a broader range of flood events and assess its generalizability across data-scarce regions and diverse climatic, topographic, and land-surface conditions.

6. Conclusions

Beyond the prevailing emphasis on refining flood-mapping algorithms, our work addresses a limitation: discrete remote sensing acquisitions can miss floodwater growth and recession between satellite revisits, leading to underestimation of MFF in existing flood inventories. To overcome this limitation, we propose a MFF extraction method based on sustained post-flood traces. The framework develops a trace-oriented identification strategy for reliable MFF reconstruction: NDWI anomalies are introduced as the primary indicator for mapping flood traces, while CCDC-based historical baselines are used to suppress seasonal variability and spatial heterogeneity. Applying the method to the once-in-a-century 2023 HRB flood event demonstrates its effectiveness. The proposed method achieves an average mapping F1-score of 81.45%, exceeding widely used standing-floodwater-based methods, whose F1-scores range only from 28.24% to 36.41%. At the catchment scale, our method reduces MFF underestimation: Compared with our estimate, MFF derived

solely from Sentinel-1 and Sentinel-2 imagery is 43.3% and 48.0% smaller, respectively. Integrating commercial SAR imagery or fusion of Sentinel-1 and Sentinel-2 slightly reduces this gap, with the mapped MFF still 34.2%, and 20.2% smaller than our estimate.

Overall, against the backdrop of increasing flood risk and substantial underestimation in remote-sensing-based flood mapping, our framework has potential for application to a broader range of flood events, providing a pathway toward more complete large-scale flood datasets.

CRedit authorship contribution statement

Yiling Lin: Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation, Conceptualization. **Xie Hu:** Writing – review & editing, Resources, Funding acquisition, Conceptualization. **Xiekang Wang:** Writing – review & editing, Resources. **Yunxia Zhang:** Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2024YFD2200900), Beijing Natural Science Foundation (8252007), and the State Key Laboratory of Hydraulics and Mountain River Engineering (SKHL2406). We thank Nan Chen for assistance with remote-sensing imagery collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2026.126783.

Data availability

Data will be made available on request.

References

- Ahmadi, P., Valadan Zoj, M.J., Mokhtarzade, M., Kardan, N., Ahmadi, P., Ghaderpour, E., 2026. TLE-FEDformer: a frequency-domain transformer framework for multi-sensor multi-temporal flood inundation mapping. *Remote Sens.* 18 (6), 895. <https://doi.org/10.3390/rs18060895>.
- Ahmadi, P., Valadan Zoj, M.J., Mokhtarzade, M., Kardan Halvaie, N., Ghaderpour, E., 2025. A capsule network framework for flood mapping integrating remote sensing fusion techniques. *Env. Res. Commun.* 7 (6), 065031. <https://doi.org/10.1088/2515-7620/ade6d0>.
- Aufaristama, M., van der Werff, H., Botha, A.E.J., van der Meijde, M., 2024. Global Landslide Finder: detecting the time and place of landslides with dense earth observation time series. *GeoHazards* 5 (3), 780–798. <https://doi.org/10.3390/geohazards5030039>.
- Bates, P.D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., et al., 2021. Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates. *Water Resour. Res.* 57 (2), e2020WR028673. <https://doi.org/10.1029/2020WR028673>.
- Baynes, E.R.C., Kincey, M.E., Warburton, J., 2023. Extreme flood sediment production and export controlled by reach-scale morphology. *Geophys. Res. Lett.* 50 (10), e2023GL103042. <https://doi.org/10.1029/2023gl103042>.
- Bereczky, M., Wieland, M., Krullikowski, C., Martinis, S., Plank, S., 2022. Sentinel-1-based water and flood Mapping: benchmarking convolutional neural networks against an operational rule-based processing chain. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 2023–2036. <https://doi.org/10.1109/jstars.2022.3152127>.
- Bofana, J., Zhang, M., Wu, B., Zeng, H., Nabil, M., Zhang, N., Elnashar, A., Tian, F., da Silva, J.M., Botão, A., Atumane, A., Mushore, T.D., Yan, N., 2022. How long did crops survive from floods caused by Cyclone Idai in Mozambique detected with multi-satellite data. *Remote Sens. Environ.* 269, 112808. <https://doi.org/10.1016/j.rse.2021.112808>.
- Chen, C.M., Hollingsworth, J., Clark, M.K., Chamlagain, D., Bista, S., Zekkos, D., et al., 2025. Erosional cascade during the 2021 Melamchi flood. *Nat. Geosci.* 18 (1), 32–36. <https://doi.org/10.1038/s41561-024-01596-x>.

- Chen, H., Liang, Q., Liang, Z., Liu, Y., Xie, S., 2019. Remote-sensing disturbance detection index to identify spatio-temporal varying flood impact on crop production. *Agric. For. Meteorol.* 269–270, 180–191. <https://doi.org/10.1016/j.agrformet.2019.02.002>.
- Chen, Z., Zhao, S., 2022. Automatic monitoring of surface water dynamics using Sentinel-1 and Sentinel-2 data with Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* 113, 103010. <https://doi.org/10.1016/j.jag.2022.103010>.
- David, C.H., Frasson, R.P.D.M., 2023. Blame the river not the rain. *Nat. Geosci.* 16 (4), 282–283. <https://doi.org/10.1038/s41561-023-01163-w>.
- Deijns, A.A.J., Michéa, D., Déprez, A., Malet, J.-P., Kervyn, F., Thiery, W., Dewitte, O., 2024. A semi-supervised multi-temporal landslide and flash flood event detection methodology for unexplored regions using massive satellite image time series. *ISPRS J. Photogramm. Remote Sens.* 215, 400–418. <https://doi.org/10.1016/j.isprsjprs.2024.07.010>.
- DeVries, B., Huang, C., Armston, J., Huang, W., Jones, J.W., Lang, M.W., 2020. Rapid and robust monitoring of flood events using Sentinel-1 and landsat data on the google earth engine. *Remote Sens. Environ.* 240, 111664. <https://doi.org/10.1016/j.rse.2020.111664>.
- Du, M., Shao, Z., Xiao, X., Zhang, J., Zhu, D., Wang, J., et al., 2026. High-precision flood change detection with lightweight SAR transformer network and context-aware attention for enriched-diverse and complex flooding scenarios. *ISPRS J. Photogramm. Remote Sens.* 231, 507–531. <https://doi.org/10.1016/j.isprsjprs.2025.11.011>.
- Esri, 2023. *World Imagery Wayback* [Basemap]. ArcGIS Living Atlas of the World. Imagery version dated June 29, 2023.
- Feng, S., Laursen, S.N., Smart, A., Sørensen, K.S., Lund, M., Grillini, F., Rieksta, J., Jiao, Y., Rinnan, R., Westergaard-Nielsen, A., 2026. Multi-source remote sensing of insect defoliation events in Abisko from point to regional scales. *Agric. For. Meteorol.* 378, 111023. <https://doi.org/10.1016/j.agrformet.2026.111023>.
- Fichtner, F., Mandery, N., Wieland, M., Groth, S., Martinis, S., Riedinger, T., 2023. Time-series analysis of Sentinel-1/2 data for flood detection using a discrete global grid system and seasonal decomposition. *Int. J. Appl. Earth Obs. Geoinf.* 119, 103329. <https://doi.org/10.1016/j.jag.2023.103329>.
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* 2 (1), 150066. <https://doi.org/10.1038/sdata.2015.66>.
- Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* 58 (3), 257–266. [https://doi.org/10.1016/S0034-4257\(96\)00067-3](https://doi.org/10.1016/S0034-4257(96)00067-3).
- Garcia-Vila, M., dos Santos Vianna, M., Harrison, M.T., Liu, K., de S. Nôia-Júnior, R., Weber, T.K.D., et al., 2025. Gaps and strategies for accurate simulation of waterlogging impacts on crop productivity. *Nat. Food* 6, 553–562. <https://doi.org/10.1038/s43016-025-01179-y>.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- Gu, X., Zhang, Q., Singh, V.P., Shi, P., 2017. Changes in magnitude and frequency of heavy precipitation across China and its potential links to summer temperature. *J. Hydrol.* 547, 718–731. <https://doi.org/10.1016/j.jhydrol.2017.02.041>.
- Hamidi, E., Peter, B.G., Munoz, D.F., Mofakhari, H., Moradkhani, H., 2023. Fast flood extent monitoring with SAR change detection using google earth engine. *IEEE Trans. Geosci. Remote Sens.* 61, 1–19. <https://doi.org/10.1109/tgrs.2023.3240097>.
- Han, J., Zhang, Z., Xu, J., Chen, Y., Jägermeyr, J., Cao, J., et al., 2024. Threat of low-frequency high-intensity floods to global cropland and crop yields. *Nat. Sustain.* 7 (8), 994–1006. <https://doi.org/10.1038/s41893-024-01375-x>.
- Lan, L., Wang, X., 2025. Large-scale flood mapping using Sentinel-1 and Sentinel-2 imagery: spatio-temporal analysis of the 23-7 Haihe basin-wide extreme flood. *J. Hydrol.* 653, 132777. <https://doi.org/10.1016/j.jhydrol.2025.132777>.
- Landuyt, L., Van Wesemael, A., Schumann, G.J.-P., Hostache, R., Verhoest, N.E.C., Van Coillie, F.M.B., 2019. Flood mapping based on synthetic aperture radar: an assessment of established approaches. *IEEE Trans. Geosci. Remote Sens.* 57 (2), 722–739. <https://doi.org/10.1109/TGRS.2018.2860054>.
- Li, J., Zhong, Y., Zhu, Q., Wang, X., Zhang, L., 2026. Unsupervised anomaly localization with high-resolution and time-series satellite imagery: from global disaster dataset to zero-shot model via SAM. *ISPRS J. Photogramm. Remote Sens.* 237, 467–491. <https://doi.org/10.1016/j.isprsjprs.2026.04.027>.
- Li, Z., Xu, S., Weng, Q., 2025. TerrainFloodSense: improving seamless flood mapping with cloudy satellite imagery via water occurrence and terrain data fusion. *Int. J. Appl. Earth Obs. Geoinf.* 144, 104855. <https://doi.org/10.1016/j.jag.2025.104855>.
- Lin, Y., Hu, X., Wang, F., Tang, Q., Zhao, Y., Li, Y., et al., 2025. Quantifying the lifespan of 3D flood structures: unlocking the potential of flood detention areas for enhanced flood control in China. *Water Resour. Res.* 61, e2025WR039992. <https://doi.org/10.1029/2025wr039992>.
- Ma, J., Gao, H., Xu, C., Qi, S., 2025. Characterization and formation mechanism of the catastrophic flash flood-debris flow hazard triggered by the July 2023 extreme rainstorm in Hantai Gully of Beijing, China. *Landslides* 22 (3), 877–893. <https://doi.org/10.1007/s10346-024-02433-3>.
- Ma, X., Jiao, J., Yang, B., Zhao, W., Ling, Q., Zhang, X., Han, J., Du, P., Chen, Y., Chen, H., 2024. Using SAR imagery to extract flash flood sediment deposition area in the northern Loess Plateau. *J. Hydrol.* 644, 132045. <https://doi.org/10.1016/j.jhydrol.2024.132045>.
- Martinis, S., Groth, S., Wieland, M., Knopp, L., Rättich, M., 2022. Towards a global seasonal and permanent reference water product from Sentinel-1/2 data for improved flood mapping. *Remote Sens. Environ.* 278, 113077. <https://doi.org/10.1016/j.rse.2022.113077>.
- Martinis, S., Kersten, J., Twele, A., 2015. A fully automated TerraSAR-X based flood service. *ISPRS J. Photogramm. Remote Sens.* 104, 203–212. <https://doi.org/10.1016/j.isprsjprs.2014.07.014>.
- McCormack, T., Campaña, J., Naughton, O., 2022. A methodology for mapping annual flood extent using multi-temporal Sentinel-1 imagery. *Remote Sens. Environ.* 282, 113273. <https://doi.org/10.1016/j.rse.2022.113273>.
- Misra, A., White, K., Nsutezo, S.F., Straka, W., Lavista, J., 2025. Mapping global floods with 10 years of satellite radar data. *Nat. Commun.* 16 (1), 5762. <https://doi.org/10.1038/s41467-025-60973-1>.
- Nemmen, A.M., Tanim, A.H., Nahian, A., Khan, S., Goharian, E., Imran, J., 2025. How extreme rainfall and failing dams unleashed the Derna flood disaster. *Nat. Commun.* 16 (1), 4191. <https://doi.org/10.1038/s41467-025-59261-9>.
- Nobre, A.D., Cuartas, L.A., Hodnett, M., Rennó, C.D., Rodrigues, G., Silveira, A., Waterloo, M.J., Saleska, S., 2011. Height above the nearest drainage: a hydrologically relevant new terrain model. *J. Hydrol.* 404 (1–2), 13–29. <https://doi.org/10.1016/j.jhydrol.2011.03.051>.
- Normand, J.C.L., Heggy, E., 2024. Assessing flash flood erosion following Storm Daniel in Libya. *Nat. Commun.* 15, 6493. <https://doi.org/10.1038/s41467-024-49699-8>.
- OPERA, 2023. *OPERA Dynamic Surface Water Extent from Harmonized Landsat Sentinel-2 (Version 1)* [Data set]. PO.DAAC. <https://doi.org/10.5067/OPDSW-PL3V1>.
- Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540 (7633), 418–422. <https://doi.org/10.1038/nature20584>.
- Portalés-Julià, E., Mateo-García, G., Gómez-Chova, L., 2025. Understanding flood detection models across Sentinel-1 and Sentinel-2 modalities and benchmark datasets. *Remote Sens. Environ.* 328, 114882. <https://doi.org/10.1016/j.rse.2025.114882>.
- Qin, Y., Yin, X., Li, Y., Xu, Q., Zhang, L., Mao, P., Jiang, X., 2025. High-Precision flood mapping from Sentinel-1 dual-polarization SAR data. *IEEE Trans. Geosci. Remote Sens.* 63, 1–15. <https://doi.org/10.1109/tgrs.2025.3557330>.
- Rahman, Md.S., Di, L., Yu, E., Lin, L., Yu, Z., 2021. Remote sensing based rapid assessment of flood crop damage using novel disaster vegetation damage index (DVDI). *Int. J. Disaster Risk Sci.* 12 (1), 90–110. <https://doi.org/10.1007/s13753-020-00305-7>.
- Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40 (8), 913–929. <https://doi.org/10.1111/ecog.02881>.
- Roth, F., Tupas, M.E., Navacchi, C., Zhao, J., Wagner, W., Bauer-Marschallinger, B., 2025. Evaluating the robustness of Bayesian flood mapping with Sentinel-1 data: a multi-event validation study. *Sci. Remote Sens.* 11, 100210. <https://doi.org/10.1016/j.srs.2025.100210>.
- Singh, K.V., Setia, R., Sahoo, S., Prasad, A., Pateriya, B., 2015. Evaluation of NDWI and MNDWI for assessment of waterlogging by integrating digital elevation model and groundwater level. *Geocarto Int.* 30 (6), 650–661. <https://doi.org/10.1080/10106049.2014.965757>.
- Tarpanelli, A., Mondini, A.C., Camici, S., 2022. Effectiveness of Sentinel-1 and Sentinel-2 for flood detection assessment in Europe. *Nat. Hazards Earth Syst. Sci.* 22 (8), 2473–2489. <https://doi.org/10.5194/nhess-22-2473-2022>.
- Tellman, B., Sullivan, J.A., Kuhn, C., Kettner, A.J., Doyle, C.S., Brakenridge, G.R., et al., 2021. Satellite imaging reveals increased proportion of population exposed to floods. *Nature* 596 (7870), 80–86. <https://doi.org/10.1038/s41586-021-03695-w>.
- Vousdoukas, M.I., Paprotny, D., Mentaschi, L., Monioudi, I.N., Feyen, L., 2026. Coastal flood impacts and lost ecosystem services along Europe's outermost regions and overseas countries and territories. *Nat. Commun.* 17 (1), 188. <https://doi.org/10.1038/s41467-025-66391-7>.
- Wagner, W., Bauer-Marschallinger, B., Roth, F., Raiger-Stachl, T., Reimer, C., McCormick, N., et al., 2026. The fully-automatic sentinel-1 global flood monitoring service: scientific challenges and future directions. *Remote Sens. Environ.* 333, 115108. <https://doi.org/10.1016/j.rse.2025.115108>.
- Wu, X., Zhang, Z., Xiong, S., Zhang, W., Tang, J., Li, Z., et al., 2023. A near-real-time flood detection method based on deep learning and SAR images. *Remote Sens.* 15 (8), 2046. <https://doi.org/10.3390/rs15082046>.
- Wu, Z., Dong, Z., Yang, K., Liu, Q., Wang, W., 2024. Floodwater extraction from UAV orthoimagery based on a transformer model. *Remote Sens.* 16 (21), 4052. <https://doi.org/10.3390/rs16214052>.
- Xu, Z., Yang, W., Zhang, X., Gu, C., Shen, L., Hu, H., 2025. Satellite-based monitoring and hazard assessment of multiple flooding types in the Haihe River Basin induced by the July 2023 extreme rainstorm. *J. Hydrol.* 663, 134243. <https://doi.org/10.1016/j.jhydrol.2025.134243>.
- Yin, J., Gao, Y., Chen, R., Yu, D., Wilby, R., Wright, N., et al., 2023. Flash floods: why are more of them devastating the world's driest regions? *Nature* 615 (7951), 212–215. <https://doi.org/10.1038/d41586-023-00626-9>.

- Zhang, Q., Gu, X., Singh, V.P., Kong, D., Chen, X., 2015. Spatiotemporal behavior of floods and droughts and their impacts on agriculture in China. *Global and Planetary Change* 131, 63–72. <https://doi.org/10.1016/j.gloplacha.2015.05.007>.
- Zhang, S., Ma, Z., Li, Z., Zhang, P., Liu, Q., Nan, Y., et al., 2021. Using CYGNSS data to map flood inundation during the 2021 extreme precipitation in Henan Province, China. *Remote Sensing* 13 (24), 5181. <https://doi.org/10.3390/rs13245181>.
- Zhang, Y., Woodcock, C.E., Chen, S., Wang, J.A., Sulla-Menashe, D., Zuo, Z., Olofsson, P., Wang, Y., Friedl, M.A., 2022. Mapping causal agents of disturbance in boreal and arctic ecosystems of North America using time series of Landsat data. *Remote Sensing of Environment*, 272, 112935. <https://doi.org/10.1016/j.rse.2022.112935>.
- Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment* 144, 152–171. <https://doi.org/10.1016/j.rse.2014.01.011>.