https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Multi-Model Comparison of Atmospheric River Flood Simulations in the Russian River Basin

Wen-Ying Wu[1], Rehenuma Lazin[1], Shiheng Duan[1], Mahjabeen Fatema Mitu[1,2], Paul Ullrich[1], Céline J. W. Bonfils[1], Hsi-Yen Ma[1], Giuliana Pallotta[1], Christopher Scott Sherman[1]

[1]Lawrence Livermore National Lab, Livermore, CA, USA
[2]Eversource Energy Center, University of Connecticut, Storrs, CT, USA

Correspondence to: Wen-Ying Wu (wu59@llnl.gov)

Abstract

Extreme atmospheric river (AR) events are a dominant driver of flood risk in Northern California. Traditional flood studies and operational forecasts often rely on a single model; however, substantial structural differences persist across hydrologic modelling frameworks due to variations in process representations. Here, we present a multi-model comparison framework to evaluate differences in streamflow and flood inundation during an AR-driven flood event in the Russian River Basin (Sonoma County, California) in late February 2019. Three physical-based modelling systems, each forced with a common observation-based precipitation dataset, are compared: (1) **Torrent**, a computationally efficient overland flow model operating directly on digital elevation models (DEMs); (2) **HEC-RAS**, implemented using a two-dimensional shallow-water (Saint-Venant) solver in rain-on-grid mode; and (3) **WRF-Hydro-HAND**, a multi-scale hydrologic routing framework that combines distributed runoff generation with terrain-based inundation mapping. These systems differ fundamentally in their treatments of runoff generation, infiltration, and flow routing, enabling an explicit comparison of model behavior across frameworks. Model performance for water level is evaluated against observations from U.S. Geological Survey stream gauges using the modified Kling–Gupta Efficiency (KGE′) and its components, including correlation, bias ratio, and variability ratio, together with diagnostics of peak timing. In addition, inundation extent is evaluated using the Landsat-derived Dynamic Surface Water Extent (DSWE) product, which provides spatially distributed observations of flood extent. Analysis of channel dynamics shows that variability and mean bias dominate model performance differences, while time correlation errors play a minor role. Results indicate that all models capture the primary inundation along the mainstem Russian River, but substantial differences emerge in tributary and floodplain regions. Torrent and HEC-RAS tend to overestimate flood extent, while WRF-Hydro-HAND tends to underestimate it. These findings highlight the importance of model structure in flood simulation and demonstrate the value of multi-model frameworks for improving understanding of flood behavior in complex watershed systems.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

1 Introduction

Inland flooding, often triggered by intense precipitation, poses a major threat to communities and critical infrastructure in the United States. According to the NOAA National Centers for Environmental Information (NCEI), flooding ranks among the costliest types of extreme events, with an average cost of approximately $4.5 billion per event (NOAA NCEI, 2024). Between 1980 and 2023, flooding in the United States caused an estimated $196 billion in damages and more than 700 fatalities (NOAA NCEI, 2024). These impacts highlight the growing need for accurate and timely flood prediction, particularly under extreme events.

Regardless of whether for operational forecasting or academic research, hydrologic studies have traditionally focused on the development and calibration of a single model within a catchment to achieve the best possible performance against observations. This approach is partly driven by computational efficiency and established practice, but it implicitly assumes that a single model structure can adequately represent the dominant hydrologic processes of a watershed. However, it is well recognized that hydrologic simulations are inherently uncertain, with uncertainties arising from input data (e.g., precipitation), model parameter estimations (e.g., Manning's coefficient), model structure (e.g., configuration, resolution, and process representation), and observational data used for evaluation and calibration (Beven, 2006; Clark et al., 2015b; Gupta et al., 2009) . Among these sources, structural uncertainty remains one of the most challenging to characterize and often contributes substantially to differences in model projections (Clark et al., 2015b).

Many studies have demonstrated that flood simulations are sensitive to structural choices, including the representation of runoff generation, routing processes, and spatial variability, with structural uncertainty often comparable in magnitude to parameter and input uncertainties (Clark et al., 2015a; Kay et al., 2009; Renard et al., 2010; Thirel et al., 2015). While uncertainties in input data and parameters can often be explored within a single modelling framework, systematic exploration of model structural differences remains limited.

To address these challenges, multi-model approaches and ensemble frameworks have increasingly been adopted as practical tools for characterizing structural uncertainty (Thirel et al., 2015; Wright et al., 2019). For example, Clark et al. (2015b) introduced the Framework for Understanding Structural Errors (FUSE), which enables systematic testing of alternative process representations. In operational forecasting, multi-model ensembles have been shown to improve robustness, particularly for events outside the calibration regime (Cloke and Pappenberger, 2009). By comparing predictions from models with differing structural assumptions, these approaches provide a useful basis for assessing variability in simulated hydrologic responses.

However, many existing studies emphasize aggregated performance metrics, with less focus on spatial evaluation of flood extent. This limitation is partly due to the scarcity of reliable, spatially distributed observational datasets for flood inundation, which has historically constrained such analyses. Recent advances in satellite remote sensing, such as the Landsat-derived Dynamic Surface Water Extent (DSWE) product, provide new opportunities for spatially distributed validation. The

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

DSWE product enables event-scale mapping of surface water from optical imagery, offering an independent dataset to evaluate modeled inundation patterns.

In this study, we present a multi-model comparison framework to examine structural uncertainty in simulations of an AR–driven flood event in the Russian River basin, Sonoma County, California. Three modelling systems—Torrent, HEC-RAS, and WRF-Hydro-HAND are evaluated. Model outputs are compared against U.S. Geological Survey (USGS) gauge observations and the Landsat-derived Dynamic Surface Water Extent Science product (DSWE). Through combined spatial and temporal evaluation, this framework provides a basis for comparing model behavior and identifying key differences in simulated flood responses.

2 Study Area and Event

Here we focus on the Russian River basin in Sonoma County, Northern California, a coastal watershed that drains into the Pacific Ocean. The study domain covers the entire Russian River basin (Figure 1), with a drainage area of approximately 3,845 km², corresponding to Hydrologic Unit Code (HUC8) 18010110. The river originates in mountainous headwaters characterized by narrow valleys and transitions downstream into broader floodplains. Elevations range up to approximately 1,200 m, with minimal snow influence. The basin experiences a Mediterranean climate with wet winters and dry summers. Most of the precipitation is from atmospheric river events in the winter. These events can produce prolonged and intense rainfall, leading to rapid runoff and elevated river stages. The lower basin, particularly near the community of Guerneville, is especially vulnerable to flooding due to its low-lying topography and limited drainage capacity. In addition, the Laguna de Santa Rosa, a large floodplain wetland complex in the lower basin, plays an important role in storing and routing floodwaters, contributing to spatially complex inundation patterns during high-flow events.

The Russian River has a history of major floods, making it a well-studied system for evaluating flood dynamics and model performance under extreme conditions (Han et al., 2019; Ma et al., 2021; Cifelli et al., 2024; Kim et al., 2025). The study focuses on an AR event that impacted Northern California in late February 2019. The event was characterized by sustained moisture transport from the Pacific Ocean, resulting in prolonged and intense precipitation across the Russian River basin. Rainfall intensified on 26–27 February, with some locations receiving more than 500 mm of precipitation over a 72-hour period, including extreme totals near Venado in Sonoma County (CW3E, 2026). The event occurred under wet antecedent conditions, as previous storms had already saturated soils in the basin, enhancing runoff generation and contributing to rapid increases in streamflow. With intense precipitation, the sharp increases in river discharge, and peak river stages were observed on 27 February evening, particularly in the lower basin near Guerneville. This event represents one of the most significant recent floods in the basin and provides a valuable case for evaluating model performance under extreme atmospheric river conditions.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

3 Materials and Methods

3.1 Model descriptions

Three hydrological modelling frameworks are compared: Torrent, HEC-RAS, and WRF-Hydro-HAND (Table 1). These three modelling frameworks were selected because each accepts precipitation directly as model forcing and generates the precipitation-to-inundation response, rather than requiring pre-computed runoff or streamflow as a boundary condition. Simulations in this study share the same atmospheric forcing, which is the NOAA Analysis of Record for Calibration (AORC) dataset with 1 km and hourly resolution (version 1.1) (Fall et al., 2023; Johnson et al., 2023). The AORC dataset is a gauge- and radar-informed gridded product designed for hydrologic applications, offering spatially consistent precipitation estimates across the United States. No event-specific calibration was performed.

Topographic information is derived from a 10 m resolution digital elevation model (DEM) obtained from the USGS (USGS, 2025). This DEM is used as the primary terrain input for both the Torrent and HEC-RAS modelling systems to represent surface elevation and flow pathways. For HEC-RAS, additional surface characteristics are incorporated using the National Land Cover Database (NLCD) 2023 (USGS, 2024). Although all final inundation products were represented on a 10 m terrain grid, only Torrent operated directly at 10 m resolution. HEC-RAS used a 100 m computational mesh with approximately 25 m channel refinement, while WRF-Hydro-HAND combined coarser land-surface and routing calculations with 10 m HAND-based inundation mapping.

Table 1. Comparison of three modelling frameworks

Models	**Torrent**	**HEC-RAS (2D Rain-on-Grid)**	**WRF-Hydro-HAND**
Primary purpose	Rapid inundation estimation for decision support	Event-based hydraulic flood simulation	Coupled hydrologic modelling framework
Core physics focus	Lagrangian rivulet approximation of overland flow	2D shallow water equations (Saint-Venant equations)	Land surface water balance, overland flow routing, and channel routing
Momentum treatment	Simplified momentum representation	Full or simplified shallow water equations	Diffusive wave routing for surface and channel flow
Infiltration representation	Not internally modeled	Parameterized loss methods (e.g., Green–Ampt, curve number)	Physically based soil moisture and infiltration via land surface model
Floodplain representation	Emergent from terrain	Explicit hydraulic connectivity	Implicit via HAND (limited lateral spread)
Typical time scale	Event-scale	Event-scale	Multi-scale (continuous simulation and event-scale)

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Implementation complexity	Low (minimal setup and rapid execution)	Moderate (mesh generation and parameter configuration required)	High (coupled system with complex setup and interpretation but automation-friendly)

3.1.1 Torrent

Torrent is an overland flow model based on a Lagrangian rivulet approach, designed to be computationally efficient compared to traditional shallow water equation solvers (Daniel et al., 2023; Daniel, 2025). The model simulates surface flow routing directly over a DEM without explicitly solving full hydrodynamic equations, instead approximating flow dynamics through simplified momentum representation and rivulet-based routing.

Due to its simplified formulation, Torrent only requires topography and dynamic forcing (precipitation). The dynamic forcing consists of precipitation from AORC, with spatially uniform Manning's roughness coefficients applied by default. Torrent computes flow depth at each time step on the native 10 m resolution of the USGS DEM. The model does not include infiltration, evapotranspiration or subsurface processes. Simulations are conducted with a time step of 1 minute with hourly output.

3.1.2 HEC-RAS

HEC-RAS (USACE, 2026; Brunner, 2016) is implemented using the two-dimensional shallow water (Saint-Venant) equations in rain-on-grid mode, which converts gridded precipitation into surface runoff and routes flow across a computational mesh using Manning-based resistance. Both channel and floodplain hydraulics are resolved through the hydrodynamic equations.

In this study, HEC-RAS version 6.7 is used. Model inputs include the 10 m resolution USGS DEM, land cover data from the NLCD 2023, and precipitation from the AORC dataset. The DEM is projected to a UTM coordinate system for model setup. Spatially distributed Manning's roughness coefficients and percent impervious area are assigned based on land cover classes using lookup tables, and infiltration is represented using the Soil Conservation Service curve number method. The computational mesh consists of a 100 m base resolution, with breaklines applied along the main channel to achieve locally refined spacing of approximately 25 m. Simulations are conducted with a computational time step of 1 minute, and model outputs are saved at hourly intervals. The model is initialized with dry conditions, and no external boundary inflows are prescribed.

Postprocessing of HEC-RAS outputs is performed using Python-based workflows applied to the output HDF5 file. HEC-RAS directly simulates water surface elevation (WSE), defined as the absolute elevation of the water surface relative to the model vertical datum. Water level/depth is derived from the simulated water surface elevation by subtracting the underlying terrain elevation. At gauge locations, water level is extracted from the corresponding computational mesh cells that contain

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

each gauge. For spatial inundation mapping, water depth is calculated by subtracting the 10 m resolution DEM from the simulated water surface elevation and then interpolated to DSWE grid for evaluation.

3.1.3 WRF-Hydro-HAND

WRF-Hydro simulations are obtained from the NOAA National Water Model (NWM) reanalysis dataset (Johnson et al., 2023). The NWM is an operational hydrologic modelling system based on WRF-Hydro (Gochis et al., 2020), providing retrospective simulations of streamflow and land surface hydrologic states across the United States. The reanalysis product is driven by 1 km AORC atmospheric forcing.

WRF-Hydro simulates runoff generation through infiltration, soil moisture dynamics, and lateral flow processes, and routes streamflow through a channel network using a diffusive wave formulation. The land surface components operate at 1 km spatial resolution, with finer routing resolution (approximately 250 m) and vector-based channel routing. The river network is derived from the National Hydrography Dataset Plus (NHDPlus). Within the study domain, stream orders range from 1 to 5, with 19,157 reaches and a total channel length of approximately 2,976 km.

In this study, streamflow outputs from the NWM reanalysis are combined with the Height Above Nearest Drainage (HAND) approach to estimate flood inundation (Liu et al., 2020, 2018; Zheng et al., 2018). The HAND method is a computationally efficient, terrain-based approach. HAND values are defined as the vertical distance between a given location and the corresponding channel elevation. The HAND dataset used in this study is derived from the 10 m resolution USGS DEM.

Streamflow is first converted to water level using predefined rating curves, which describe the relationship between discharge and stage height. The HAND dataset provides these rating curves as lookup tables at discrete water levels, and linear extrapolation is used to estimate continuous stage values (Liu et al., 2020). Flood depth at each location is then calculated by comparing the stage height with the corresponding HAND value. If the stage height exceeds the HAND value, the location is considered flooded, and water depth is defined as their difference; otherwise, the location is considered dry. This approach provides an efficient means of estimating spatial flood extent without explicitly solving full hydrodynamic equations.

3.2 Observational Data for Model Evaluation

3.2.1 Streamflow gauge water level

Temporal dynamics in channel water level are evaluated using in-situ stream gauge data obtained from the USGS (USGS, 2020a). A total of 28 gauges across the basin provide instantaneous gauge height observations at 15-minute intervals, which are aggregated to hourly resolution to match the model outputs.

USGS gage height (stage), h, is referenced to a site-specific gage datum rather than the local channel bed, and therefore cannot be directly compared with model-simulated water level. To obtain a water level (L) referenced to the channel

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

bed, we define the gage height at zero flow, GZF, as the stage corresponding to zero discharge at a given gauge. Water level (L) is then computed as

$$L = h - GZF \quad (1)$$

GZF was obtained from the zero-discharge parameter of the USGS rating curve.

For gauges where a GZF could not be determined, water level was instead estimated using an independent, DEM-based reference. Gage height was first converted to an absolute water surface elevation (WSE, referenced to NAVD88) by adding the USGS gage datum elevation, D:

$$WSE = D + h \quad (2)$$

Water level was then computed as the difference between WSE and the terrain elevation, Z_DEM, extracted from a 10 m DEM at the adjusted gauge location (see Section 3.2.2):

$$L = WSE - z_{dem} = D + h - z_{dem} \quad (3)$$

When both methods were available for a given gauge, the reference yielding fewer non-physical (negative) values during the analysis period was selected, as negative water level (L) is not physically meaningful (listed in Table A1). Model performance was subsequently evaluated against these processed USGS water level time series. These observation-based water levels are used to assess model performance in terms of timing, peak magnitude, and overall temporal dynamics.

3.2.2 Gauge location adjustment for model evaluation

When comparing Torrent outputs at 10 m resolution with USGS stream gauge observations, a systematic spatial mismatch was identified. Many gauge coordinates corresponded to zero or near-zero modelled water depths, while higher water depths were found in adjacent grid cells. The average offset between gauge locations and model-predicted peak depths was approximately 3.7 grid cells. Further investigation indicated that this mismatch reflects the physical placement of gauges rather than coordinate or projection errors. USGS metadata suggests that reported gauge coordinates often correspond to accessible locations, such as structures on riverbanks or bridges, rather than the channel center where peak flow depths usually occur. This interpretation is supported by visual inspection of satellite imagery, which shows that many gauges are positioned along channel margins.

To address this mismatch, gauge locations were adjusted to better align with the river network. We first used the Stream gage Watershed InforMation (SWIM) dataset (Hayes et al., 2023), which provides USGS gauge locations referenced to the NHDPlus v2.1 river network. In the SWIM dataset, gauge locations are mapped to the centerline of river reaches using automated linear referencing with manual quality control, making them more representative of channel flow conditions. Within the Russian River basin, 12 of the 28 gauges are available in the SWIM dataset, with reported snap distances of 25.7 m on average. However, SWIM does not include all gauges and is based on the coarser NHDPlus v2.1 network. Therefore, we performed a nearest-neighbor snapping procedure using the higher-resolution NHDPlus High Resolution (HR) river network (USGS, 2022). Each gauge was mapped to the nearest river reach based on spatial proximity, ensuring alignment with the

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

channel centerline at finer spatial resolution. All adjusted gauge locations were subject to quality control by comparing original USGS coordinates, SWIM-based locations, and NHDPlus HR-derived positions (Appendix A). This procedure ensures consistent and physically meaningful alignment between observation points and model grid cells.

3.2.3 Landsat-derived flood extent

Spatial flood extent is evaluated using the Landsat-derived DSWE product (USGS, 2020b; Jones, 2019). The DSWE dataset provides a snapshot of surface water from optical satellite imagery, enabling spatially distributed assessment of inundation patterns across the basin. The product has a spatial resolution of 30 m, allowing fine-scale comparison with model simulations. The DSWE data were postprocessed to generate a consistent reference flood map for model evaluation. We use the Landsat tile acquired on 28 February 2019 at 18:52 UTC (DSWE Tile Identifier: LC08_CU_001008_20190228_20210503_02_DSWE), including the interpreted water classification (INTR) and associated quality mask layers.

The dataset is first clipped to the study domain. Flooded pixels are defined as INTR classes 1 (high confidence) and 2 (moderate confidence) under cloud-free conditions, while pixels affected by cloud or shadow are excluded from analysis. The resulting reference map is a binary raster of flooded and non-flooded areas. Compared to traditional gauge-based validation, DSWE provides additional insight into floodplain extent, which is critical for evaluating inundation performance. Model outputs are extracted at 19:00 UTC and resampled to the DSWE 30 m grid and projection to ensure consistent spatial comparison. The available DSWE coverage captures inundation patterns in the lower basin, including areas west of Santa Rosa and along the main channel. However, spatial coverage is limited and does not fully include regions such as Sebastopol and Guerneville, where significant flooding occurred during the event.

3.3 Evaluation Metrics

The evaluation period spans 72 hours from 26 February 2019 16:00 UTC (08:00 local time) to 1 March 2019 15:00 UTC (07:00 local time), capturing the onset, peak, and recession of the flood event. For temporal dynamics, simulated water level is compared against USGS-based water level using the modified Kling–Gupta Efficiency (KGE′; Gupta et al., 2009; Kling et al., 2012), which provides a measure of model skill by combining correlation, bias, and variability components. KGE′ is defined as:

$$KGE' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \quad (4)$$

where

$$\beta = \frac{\mu_s}{\mu_o}, \quad (5)$$

and

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

$$\gamma = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \tag{6}$$

r represents the linear correlation between simulated and observed time series, β is the bias ratio mean of simulations divided by observations), and γ is the variability ratio (normalized standard deviations of simulations divided by observations). μ_s and μ_o are the mathematical means of the simulated and observed data. σ_s and σ_o are the standard deviations of the simulated and observed data. This decomposition allows evaluation of different aspects of model performance, including temporal correlation, magnitude bias, and variability bias. A KGE′ value of 1 indicates perfect agreement between simulated and observed water level, whereas a value of -0.41 corresponds to performance no better than simply using the observed mean as a predictor (Knoben et al., 2019); values below -0.41 indicate performance worse than this benchmark.

In addition to KGE′, event-based metrics are used to assess peak flow characteristics. Peak magnitude bias is quantified as the ratio of simulated to observed peak water level, noting that the peaks do not necessarily occur at the same time. Peak timing error is defined as the difference in time between simulated and observed peak occurrences. These metrics provide additional insight into model performance during extreme events.

For spatial evaluation, simulated flood extent is compared with the Landsat-derived DSWE product. Model outputs are first resampled to the 30 m resolution of the DSWE dataset. We classified modeled inundation as “flooded” when simulated water depth exceeded 0.1 m and “non-flooded” otherwise. DSWE classes 1 and 2 were treated as observed inundation, while cloud/shadow-affected pixels were excluded from the comparison domain. The evaluation is based on a confusion-matrix framework, which is commonly used in remote sensing classification and binary detection analysis (Congalton, 1991; Fawcett, 2006). For each valid pixel, agreement is classified as true positive (TP, correctly simulated flooded), false positive (FP, overestimated flooded), false negative (FN, missed flooded), or true negative (TN, correctly simulated non-flooded).

Based on the confusion matrix, we compute Intersection over Union (IoU), recall, and bias to quantify model performance. IoU measures the overlap between modelled and observed flooded areas and is defined as:

$$IoU = \frac{TP}{TP+FP+FN}, \tag{7}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Recall (also referred to as probability of detection) measures the fraction of observed flooded pixels correctly identified by the model:

$$Recall = \frac{TP}{TP+FN}, \tag{8}$$

Bias quantifies the tendency of the model to overpredict or underpredict flood extent and is defined as:

$$Bias = \frac{TP+FP}{TP+FN}, \tag{9}$$

A bias greater than 1 indicates overprediction of inundation extent, while a value less than 1 indicates underprediction.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

4. Results

To illustrate overall model performance, Figure 1 presents the maximum simulated flood depth (threshold at >0.1 m) during the event at models' 10 m resolutions. The February 2019 AR event produced widespread flooding across the Russian River basin, with the most extensive inundation occurring in the lower basin along the mainstem Russian River and adjacent floodplains between Healdsburg and Guerneville, and the Laguna de Santa Rosa. In contrast, flooding in the upper basin is largely confined to narrow valleys and channel corridors.

All three modelling systems capture the general spatial pattern of inundation; however, notable differences emerge in the extent and distribution of flooding, particularly in low-lying floodplain regions. For example, in the Laguna de Santa Rosa, WRF-Hydro-HAND produces a more limited inundation extent compared to Torrent and HEC-RAS. At the basin scale, HEC-RAS yields the largest inundated area (17.2%), followed closely by Torrent (16.9%), while WRF-Hydro-HAND (7.1%) produces a smaller and more spatially fragmented flood extent. HEC-RAS simulations exhibit more spatially continuous inundation patterns, including isolated shallow flooded areas, suggesting sensitivity to local terrain and rainfall-driven surface processes. Torrent produces more structured inundation patterns in flat and urban regions, where features such as road networks are more clearly resolved. Zoomed-in views near Guerneville (Figures 1d–f) further highlight differences among the models, particularly in floodplain connectivity and local inundation patterns. Overall, the spatial patterns simulated by Torrent and HEC-RAS show closer agreement with reported impacts and evacuation records in this region. In contrast, WRF-Hydro-HAND exhibits more discontinuous inundation patterns associated with its reach-based representation of flood extent and does not include inundation in lake areas due to the methodology. These differences reflect varying model representations of runoff generation, flow routing, and floodplain connectivity.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

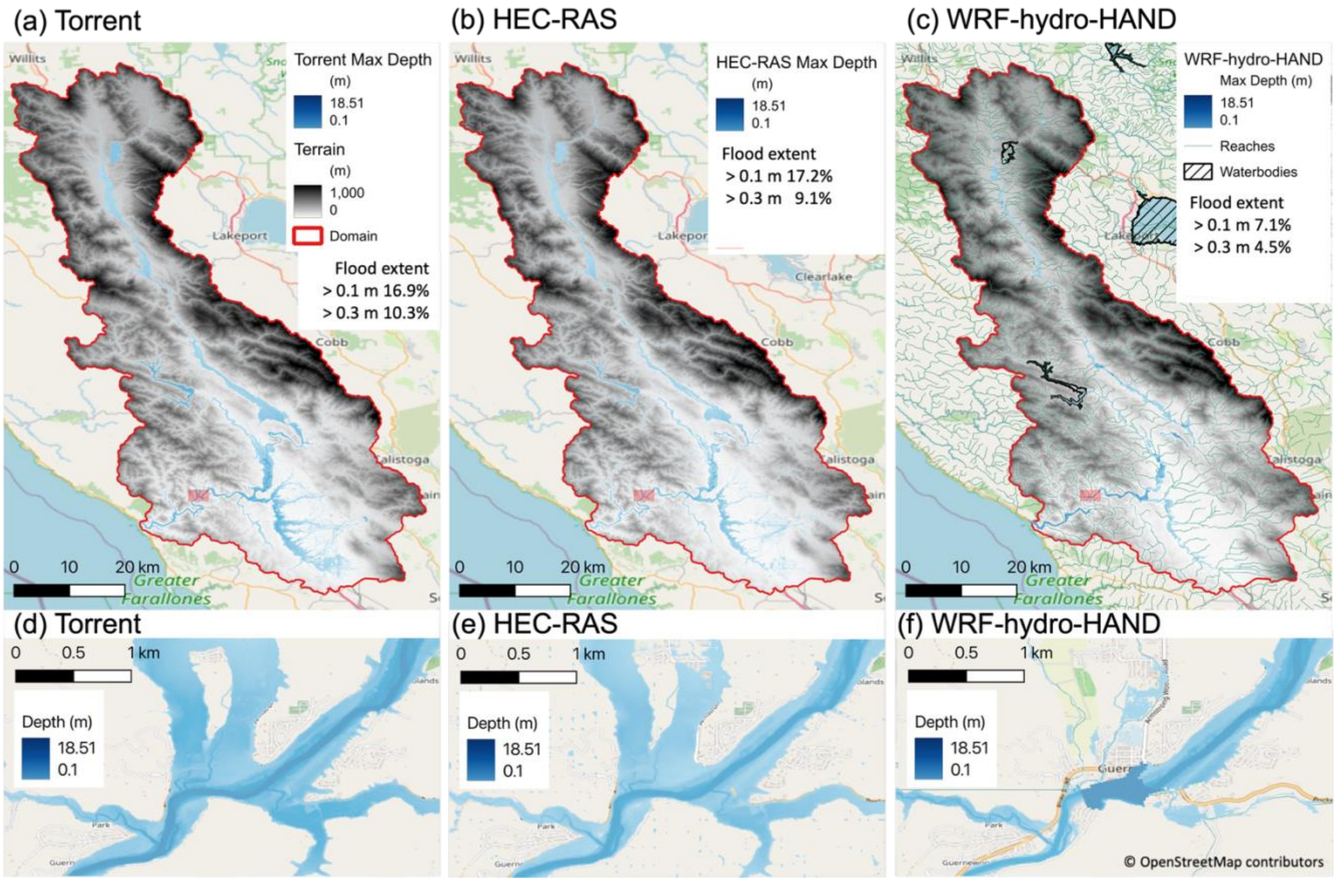


Figure 1: Maximum simulated flood depth (>0.1 m) for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND with NHDPlus reaches marked. All maps shown at their native 10 m resolution. Red shaded boxes indicate regions of Guerneville shown in panels (d–f).

To further quantify model performance, we evaluate channel dynamics using the KGE′ and its components at 28 USGS gauges over the 72-hour analysis period. All models are evaluated without calibration to ensure a consistent comparison framework. It is important to note that the three modelling systems represent channel processes differently. Torrent routes surface water directly across the DEM without a predefined channel network, allowing flow paths to emerge from topography. In HEC-RAS rain-on-grid simulations, both main channel and floodplain flows are governed by the same hydrodynamic equations, but differences in mesh resolution, terrain representation, and roughness parameters result in distinct flow behavior along the main channel compared to surrounding floodplain areas. In contrast, WRF-Hydro-HAND relies on a predefined river network with reach-based routing. As a result, KGE′ reflects performance in simulated water level at gauge locations rather than strictly channel routing process.

Spatial distributions of KGE′ provide a complementary perspective to the temporal dynamics, highlighting where model skill is spatially consistent or location specific. A KGE′ value of 1 indicates perfect agreement between simulated and

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

observed time series, while a value of -0.41 indicates performance comparable to using the observed mean (Knoben et al., 2019), and values below this threshold indicate performance worse than that benchmark. The spatial patterns of KGE′ reveal distinct differences among the models (Figure 2-3; Appendix B). Torrent shows relatively higher skill along portions of the mainstem Russian River in the middle to lower basin (Figure 2a). HEC-RAS generally exhibits stronger performance along the main channel compared to tributaries (Figure 2b), reflecting its refined mesh and improved representation of channel geometry. WRF-Hydro-HAND shows relatively better performance in the upper and middle basins, including Dry Creek, but performance decreases toward the lower basin and river outlet (Figure 2c).

Across all models (Figures 2a–c), lower skill is generally observed in tributaries such as Big Sulphur Creek and in the Laguna de Santa Rosa region, where KGE′ values remain low across multiple gauges. In the Laguna de Santa Rosa, these gauges are associated with small drainage areas and complex wetland or floodplain dynamics, which are challenging to represent across all modelling systems. In contrast, all models show relatively consistent performance near Healdsburg where agreement between simulated and observed water level is higher compared to other locations. Notably, model behaviour is not always consistent, underscoring that no single model consistently outperforms the others across all locations.

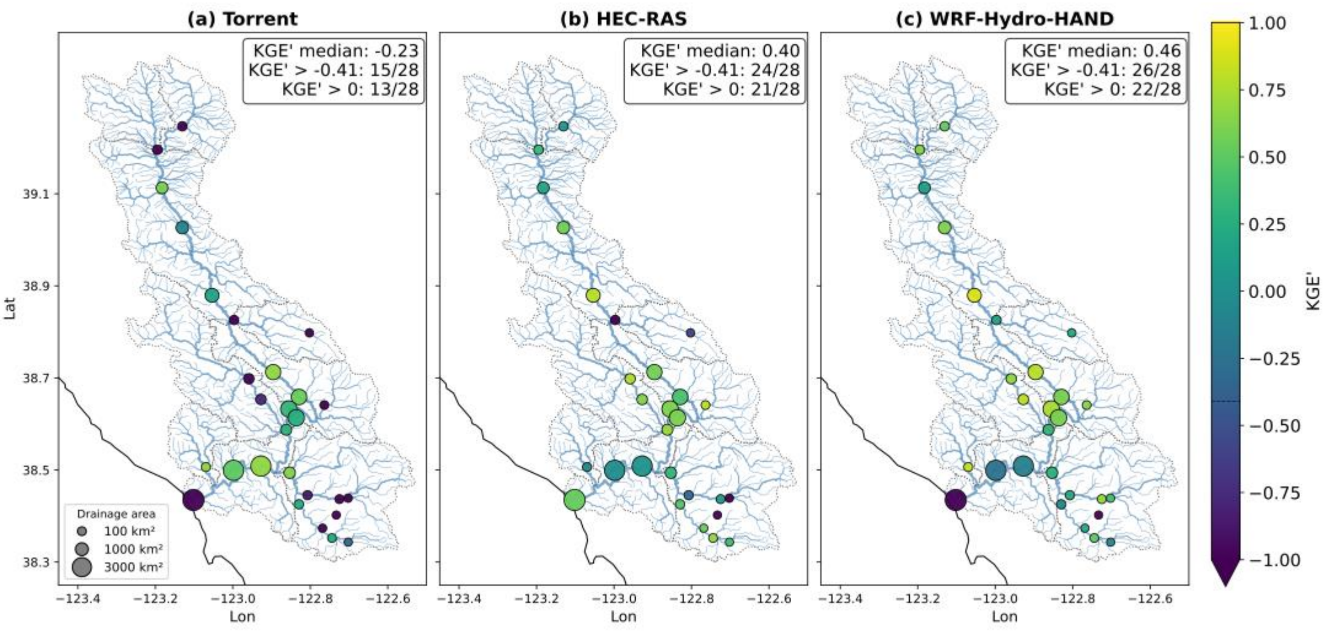


Figure 2. Spatial distribution of modified Kling–Gupta Efficiency (KGE′) for simulated water level at 28 USGS gauges over the 72-hour event period for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. KGE′ values are computed from hourly time series and quantify agreement between simulated and observed water level at each gauge. Marker size is proportional to the upstream drainage area reported by USGS.

To better understand the sources of model error, we examine the dominant (weakest) components contributing to KGE′ at each gauge (Figure 3). By identifying the weakest component among correlation, bias, and variability, this analysis provides insight into the primary factors limiting model performance.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

For Torrent (Figure 3a), variability (purple) is the most common limiting factor (16 of 28 gauges), with median variability ratios greater than 1 (γ = 1.33), indicating excessive fluctuations in simulated water level. Bias (blue) is also present at several locations in the lower basin, suggesting systematic over- or underestimation at these sites, consistent with its elevated median bias ratio (β = 1.27). Correlation errors are comparatively rare (1 gauge). HEC-RAS (Figure 3b) shows a different spatial distribution of error sources. Variability remains dominant (19 of 28 gauges). Bias-related errors are also observed at 7 gauges. Correlation errors (orange) appear at two mainstem locations. WRF-Hydro-HAND (Figure 3c) also shows variability as the dominant error source (13 gauges), followed by bias (10 gauges) , but with a comparatively higher number of gauges limited by correlation errors (5 of 28) than the other two models. Across all models, variability more prevalent than bias and then correlation errors, suggesting that variability contributes more strongly to model performance.

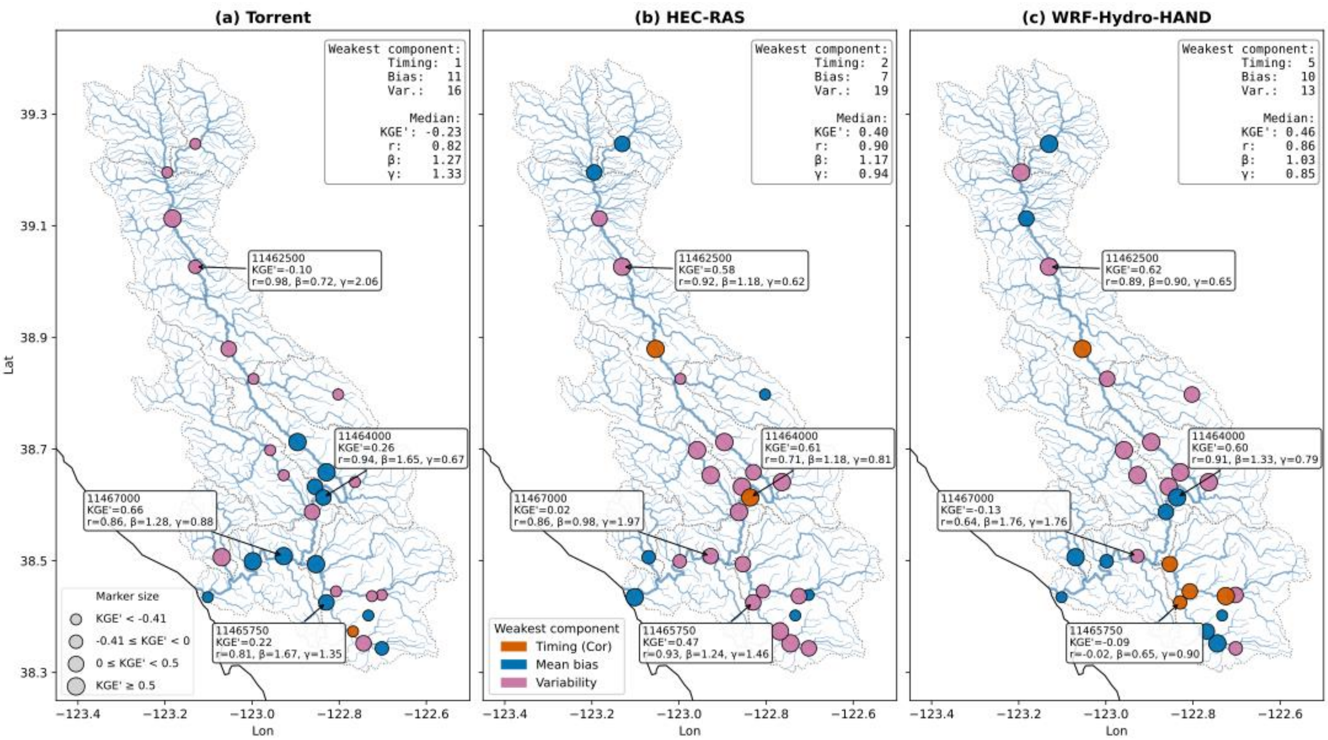


Figure 3. Weakest component of KGE′ at each USGS gauge for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Colors indicate the primary source of model error: red denotes correlation, blue denotes mean bias, and green denotes variability. Marker size is proportional to the corresponding KGE′ value, with larger markers indicating better overall model performance. Values of each component are showed in Figure B1-B3.

To further illustrate these differences, we examine time series at selected representative gauges (location highlighted in Figure 3, and timeseries in Figure 4). At the upstream site near Hopland (ID: 11462500; Figure 4a), variability is the weakest component in all three models. HEC-RAS captures the overall timing and magnitude but shows discrepancies during the initial hours of the simulation, which may be associated with model initialization. WRF-Hydro-HAND exhibits relatively damped variability. Torrent underperforms at this location due to high fluctuations (γ = 2.06), indicating sensitivity to high-resolution routing.

At middle mainstream near Healdsburg (11464000; Figure 4b), models generally produce better skill, but the dominant source of error differs across models. Bias remains the weakest component for Torrent (β = 1.65) and WRF-Hydro-HAND (β = 1.33), both of which systematically overestimate water level and its peak. HEC-RAS, in contrast, is limited primarily by correlation (r = 0.71), characterized by a faster rising limb and delayed recession

In the Laguna de Santa Rosa region (11465750; Figure 4c), the dominant source of error varies substantially by model. Torrent is limited by bias (β = 1.67) and reflecting a sustained overestimation of the elevated plateau water level rather than underestimation. HEC-RAS is limited by variability (γ = 1.46), broadly capturing the peak magnitude and timing while overrepresenting overall fluctuation relative to smoother observed response. WRF-Hydro-HAND shows a near-zero correlation (r = -0.02), the weakest value among four examined sites, reflecting an early decline in simulated water level while observations continue to rise. Observations indicate a relatively flat hydrograph at elevated water level, suggesting sustained storage and attenuation processes, indicating distinct hydrologic behavior compared to upstream channelized reaches. These conditions are characteristic of complex wetland and floodplain systems, which are challenging to represent across all modelling frameworks.

At the downstream site near Guerneville (11467000; Figure 4d), a well-documented major flood location, bias is the dominant source of error for Torrent, which overestimates water level, consistent with a modest overestimation and without clear recession. HEC-RAS exhibits variability-related errors but captures the overall event magnitude more consistently. WRF-Hydro-HAND shows comparably large deviations in both bias and variability (β = 1.76, γ = 1.76), reflecting a large, transient overestimation during the rising limb that substantially exceeds the observed peak. Importantly, all models exceed the “major” flood threshold defined by the California–Nevada River Forecast Center, indicating that the severity of the event is broadly reproduced despite differences in magnitude and timing. This consistency at a critical flood-prone location suggests that, while model discrepancies exist, the primary hazard signal is represented across modelling frameworks.

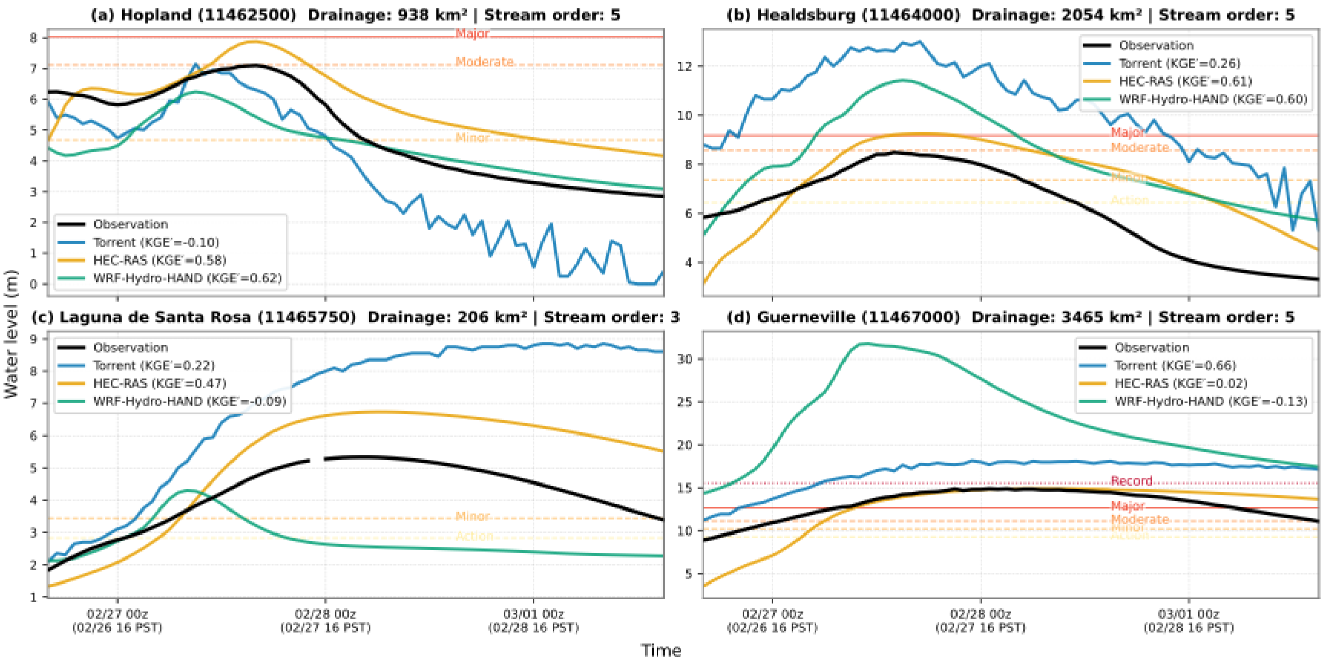


Figure 4. Time series of simulated and observed water level at four representative USGS gauges: (a) 11462500 (upper basin near Hopland), (b) 11464000 (middle basin near Healdsburg), (c) 11465750 (Laguna de Santa Rosa near Sebastopol), and (d) 11467000 (lower basin near Guerneville). Colored lines represent model simulations (Torrent, HEC-RAS, and WRF-Hydro-HAND), and black lines denote observations. Horizontal dashed lines indicate thresholds converted from flood gauge heights from the California–Nevada River Forecast Center.

We also compare peak timing errors which provide additional insight into model performance (Figure 5). Torrent tends to produce earlier peaks in the upper basin. WRF-Hydro-HAND generally exhibits earlier peak timing across many gauges, particularly in the lower basin, indicating faster runoff response compared to observations. Spatially, peak timing errors are more pronounced in wetland and floodplain regions, where flow dynamics are complex. In these areas, Torrent and HEC-RAS tend to produce later peak timing, while WRF-Hydro-HAND more frequently produces earlier peaks. These contrasting patterns highlight differences in how models represent flow routing and floodplain storage across the basin.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

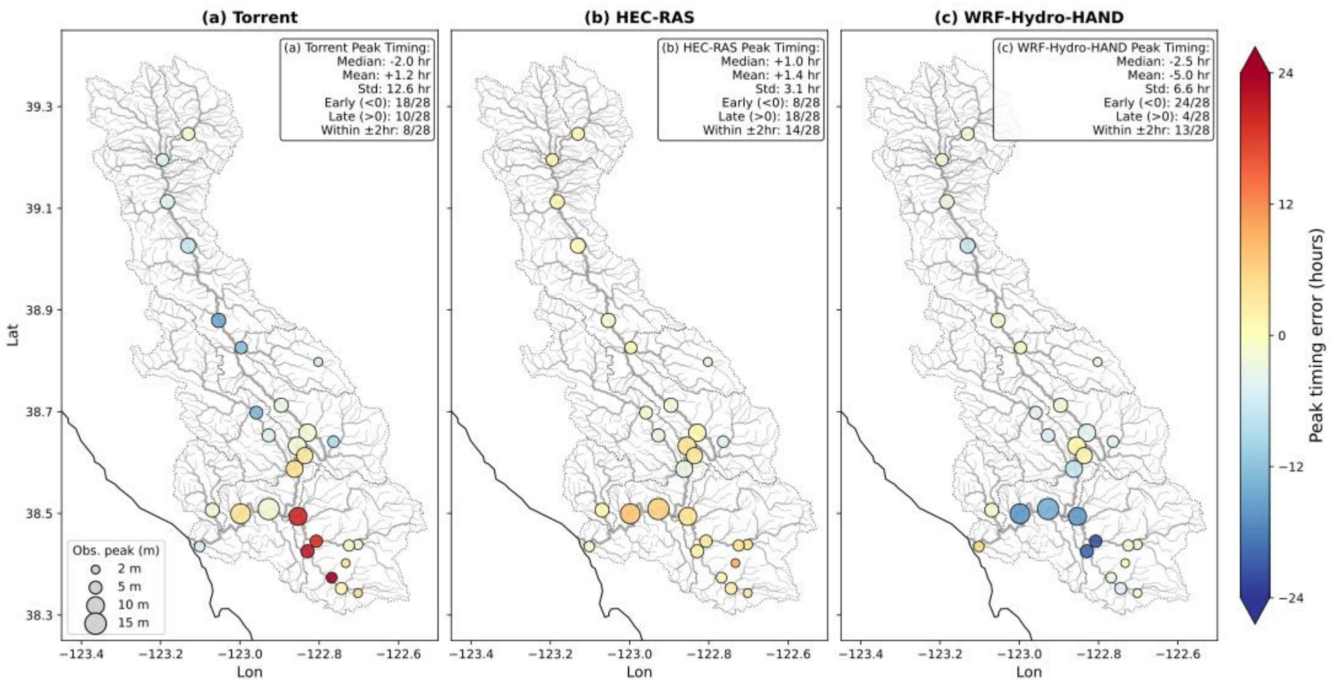


Figure 5. Peak timing error (hours) at USGS gauges for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Marker color indicates timing differences relative to observations (blue: earlier; red: later), and marker size is proportional to observed peak water level.

While time series analysis at stream gauges provides valuable insight into water level dynamics, these observations are limited to discrete channel locations and do not capture the spatial extent of flooding across the floodplain. In the absence of distributed, ground-based flood depth observations, stream gauge records remain the most reliable source of continuous temporal information; however, a single Landsat-derived Dynamic Surface Water Extent (DSWE) scene offers a complementary, spatially distributed snapshot of inundation at one point in time. We use this satellite-derived product to evaluate how well each model reproduces the observed flood extent, providing a spatial perspective that the gauge network alone cannot offer. Model-simulated flood depth is first interpolated to the 30 m DSWE resolution and then converted to a binary wet/dry classification using a threshold of 0.1 m. The Landsat acquisition (28 February 2019, 18:52 UTC; 10:52 local time) occurs after the peak stage of the flood event. At the Guerneville gauge, the Russian River reached approximately 47.55 ft around 19:15 local time on 27 February 2019, while gauge height remained elevated at about 44.77 ft at the time of satellite overpass. This indicates that the DSWE observation captures the recession phase of the flood, when inundation remained extensive but had begun to recede. Figure 6 compares model-simulated flood extent with the DSWE reference using a confusion-matrix framework. The satellite image primarily covers the lower basin west of Santa Rosa, where the mainstem Russian River turns sharply westward. The observed inundation includes portions of the Sebastopol area but does not extend far enough downstream to capture the Guerneville region, where peak flooding impacts were reported.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Across all models, the main river corridor is well captured, as indicated by extensive true positive regions (blue). However, substantial differences are observed in surrounding floodplain areas. Torrent and HEC-RAS both show widespread false positives (red), particularly in low-lying regions, indicating overestimation of flood extent beyond observed inundation. This overprediction is more spatially extensive in Torrent, while HEC-RAS produces a more continuous and spatially coherent inundation pattern along the main channel and adjacent floodplain. In contrast, WRF-Hydro-HAND exhibits a higher proportion of false negatives (yellow), especially along floodplain margins, indicating underestimation of inundation extent and a tendency to confine flooding closer to the channel network. These results highlight a systematic contrast across models: Torrent and HEC-RAS tend to overestimate flood extent, while WRF-Hydro-HAND tends to underestimate it. Despite these differences, all models capture the primary inundation along the mainstem, suggesting that large-scale flood patterns are consistently represented, while discrepancies are concentrated in floodplain and low-relief areas.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

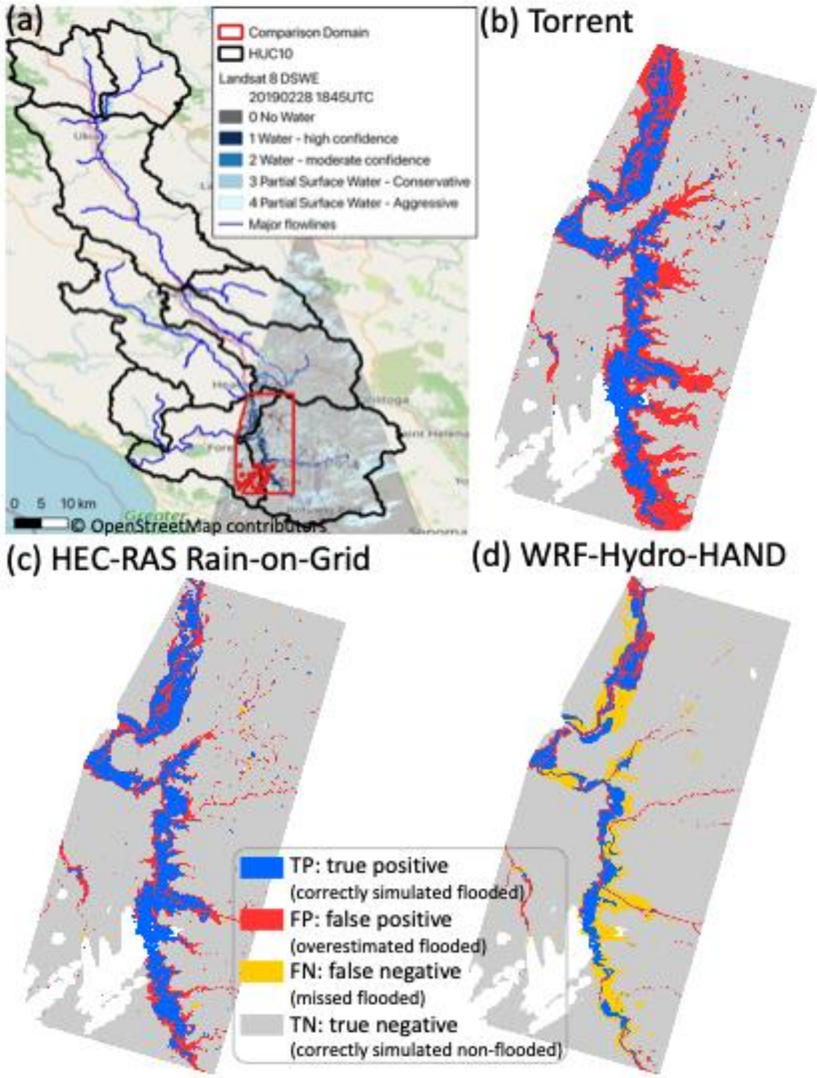


Figure 6. Comparison of model-simulated flood extent with Landsat-derived Dynamic Surface Water Extent (DSWE): (a) location of the satellite coverage within the study domain (highlighted by the red box), (b) Torrent, (c) HEC-RAS (rain-on-grid), and (d) WRF-Hydro-HAND. Model outputs are converted to binary inundation using a 0.1 m depth threshold and compared with DSWE observations at 30 m resolution. Colors indicate classification outcomes: true positives (blue), false positives (red), false negatives (yellow), and true negatives (gray). The DSWE observation corresponds to 28 February 2019, 18:52 UTC (10:52 local time), and is compared with the nearest model output at 19:00 UTC.

In terms of quantitative performance (Table 2), HEC-RAS demonstrates the strongest overall skill, achieving the highest IoU (0.604) while maintaining high recall (0.977) and moderate bias (1.59), indicating a balanced representation of inundation extent. Torrent achieves an extremely high recall (0.994), capturing nearly all observed flooded areas, but at the cost of a substantially elevated false positive rate (16.6%) and bias (2.34), indicating systematic overestimation. In contrast, WRF-Hydro-HAND shows the lowest false positive rate (3.4%) and highest true negative rate (84.3%), but markedly reduced

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

recall (0.461), missing a significant portion of observed flooding and resulting in the lowest IoU (0.361). Overall, these results indicate that Torrent tends to overestimate inundation extent, WRF-Hydro-HAND tends to underestimate it, and HEC-RAS provides the most balanced representation.

These contrasting behaviors reflect fundamental differences in how each modelling system represents runoff generation, flow routing, and floodplain connectivity.

Table 2. Summary of confusion metrics for simulated flood extent compared with DSWE observations. Metrics include true positive rate (TP), false positive rate (FP), false negative rate (FN), true negative rate (TN), Intersection over Union (IoU), recall, and bias. Metrics are computed within the valid comparison domain (red highlighted area in Figure 6a).

	Torrent	HEC-RAS	WRF-Hydro-HAND
TP	12.2%	12.0%	5.7%
FP	16.6%	7.6%	3.4%
FN	0.1%	0.3%	6.6%
TN	71.1%	80.1%	84.3%
IoU	0.42	0.60	0.36
Recall	0.99	0.98	0.46
Bias	2.34	1.59	0.74

5. Discussion

This study compares three flood modelling frameworks with substantially different process representations and computational characteristics during the February 2019 atmospheric river flood in the Russian River basin. Under common precipitation forcing and without model-specific calibration, all three frameworks reproduce the major flood signal, but their performance varies substantially by locations and metrics. Agreement is generally stronger along the middle mainstem and weaker in tributaries and wetland-dominated regions, with no model consistently outperforming the others at all gauges. Spatial evaluation similarly shows broad agreement in major inundated areas but substantial differences in floodplain extent. Because all three frameworks are directly forced by precipitation and evaluated without calibration, the comparison emphasizes differences arising from model structure and process representation rather than model-specific parameter tuning.

The three approaches differ substantially in computational and implementation complexity (Table 1). Torrent provides a relatively simple and computationally efficient framework, requiring minimal setup and enabling rapid overland-flow simulation directly from terrain and precipitation inputs. HEC-RAS requires more detailed configuration, including mesh generation and parameter specification, and typically involves greater manual setup. WRF-Hydro, particularly when implemented through the National Water Model (NWM), provides a comprehensive hydrologic framework that is well suited

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

to script-based and high-performance computing environments, but its coupled land-surface and routing components increase setup and interpretation complexity. These differences illustrate an important tradeoff between computational efficiency, process representation, and the level of hydraulic detail required for flood simulation.

The contrasting inundation patterns can be interpreted through differences in how the three frameworks represent pluvial and fluvial processes (Table 1). Torrent routes precipitation-driven surface water directly across the terrain without explicit infiltration or a predefined channel network, producing a pluvial-dominated response and relatively widespread inundation. WRF-Hydro-HAND follows a more fluvial-oriented approach, in which routed channel flow is converted to stage and mapped onto the surrounding terrain using HAND, tending to constrain inundation closer to the river network. HEC-RAS occupies an intermediate position by solving the two-dimensional shallow-water equations for both rainfall-driven surface flow and channel–floodplain hydraulics, allowing coupled pluvial–fluvial interactions. These structural differences are consistent with the tendency for Torrent to produce more extensive inundation, WRF-Hydro-HAND to produce more confined inundation, and HEC-RAS to provide a more balanced representation of observed flood extent. Similar contrasts between reduced-complexity and hydrodynamic inundation approaches have been reported by Afshari et al. (2018), who found that HEC-RAS 2D generally resulted in larger inundated area than low-complexity HAND at the downstream regions, while local differences were strongly influenced by terrain representation and hydraulic controls.

Analysis of KGE′ components provide further insight into the sources of model disagreement. Variability and bias are more frequently identified as the weakest components than correlation across all three frameworks. Although KGE′ weights the three components equally, the common precipitation forcing likely contributes to broadly consistent temporal responses, whereas differences in runoff generation, infiltration, routing, channel representation, and floodplain storage more strongly affect the magnitude and variability of simulated water levels. This decomposition is useful because similar overall KGE′ values can result from substantially different sources of model error (Gupta et al., 2009; Kling et al., 2012; Knoben et al., 2019). The spatial variability in performance is also consistent with regional evaluations of the NWM. For example, Kim et al. (2025) found substantial spatial differences in short-term NWM streamflow skill across the broader San Francisco Bay region, with performance varying with flow regime and watershed characteristics, suggesting that the heterogeneous performance observed here is not unique to a single Russian River flood event.

Because correlation reflects overall temporal agreement rather than peak timing specifically, peak timing is evaluated separately to examine differences in flood-wave propagation. Torrent tends to produce earlier peaks in smaller upstream basins, consistent with rapid conversion of precipitation to surface flow and limited storage or routing delay. The spatial pattern of Torrent timing errors may also reflect its use of a spatially uniform Manning's roughness coefficient rather than a distributed parameterization informed by land cover, as previous sensitivity studies have demonstrated the importance of Manning's roughness for flood-wave timing (Ye et al., 2018). HEC-RAS and WRF-Hydro generally show better agreement with observed peak timing in these upstream regions, although systematic timing differences remain at individual gauges.

Low-relief floodplain environments present a different challenge. In the Laguna de Santa Rosa, complex floodplain and wetland storage, backwater influences, and potential local hydraulic controls contribute to prolonged and attenuated water-

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

level responses that are difficult for all three frameworks to reproduce. Models that do not explicitly represent these storage and connectivity processes may therefore exhibit systematic differences in both magnitude and timing. WRF-Hydro-HAND generally produces earlier peaks in this region, whereas HEC-RAS and Torrent tend toward delayed peaks, illustrating how different representations of floodplain storage and routing influence event timing.

Spatial evaluation against the Landsat-derived DSWE product further distinguishes the three modelling approaches. HEC-RAS provides the most balanced agreement with observed inundation and shows particularly strong agreement in the Laguna de Santa Rosa, where agricultural, wetland, and developed areas form a heterogeneous low-relief floodplain. This result is consistent with Nikrou et al. (2026), who found that HEC-RAS performed comparatively well in urban and agricultural environments relative to other inundation models, suggesting potential advantages of explicit two-dimensional hydraulic representation in complex floodplain environments. In contrast, WRF-Hydro-HAND produces substantially less floodplain inundation than Torrent and HEC-RAS.

Several characteristics of the DSWE reference should nevertheless be considered when interpreting the spatial metrics. The DSWE classification contains multiple surface-water categories with varying confidence levels. We use only high- and moderate-confidence water classes (classes 1 and 2), providing a conservative definition of inundation but potentially omitting narrow tributaries, shallow water, and partially inundated or vegetated pixels. Consequently, some areas classified as model false positives may reflect limitations of the observational classification rather than true model overestimation. DSWE also includes pre-existing and permanent surface water because a pre-event water mask is not removed in this analysis. However, the pronounced expansion of surface water between pre-event and event-time Landsat imagery indicates that the February 2019 flood contributed substantially to the observed spatial extent (NASA, 2019).

An additional consideration is that the Landsat overpass represents only a single snapshot during the recession phase of the flood. Nikrou et al. (2026) showed that relative performance among flood inundation models can vary between rising and falling flood stages and with the observational benchmark used for evaluation. The bias reported here should therefore be interpreted as measures of inundation performance at the time of the Landsat overpass rather than as event-wide rankings of model performance. Despite these limitations, DSWE provides an independent spatial observation that complements the point-based gauge evaluation and enables assessment of model behavior across major floodplain regions.

Several additional methodological limitations should be considered. First, gauge-based evaluation introduces uncertainty in matching point observations to high-resolution simulations in both the vertical and horizontal dimensions. Vertically, USGS gage height is referenced to a site-specific gage datum rather than the local channel bed and therefore is not directly comparable with model-simulated water levels. We convert observed gage height to water level relative to a local zero-flow or terrain-based reference, as described in Section 3.2.1; however, uncertainty in the estimated channel-bed reference remains and may contribute to model–observation differences. Horizontally, reported USGS gauge coordinates may correspond to accessible bank or bridge locations rather than the channel center. This issue is particularly important for Torrent, where water levels are simulated directly on the 10 m DEM and small spatial offsets can produce substantial differences in local values. Gauge locations are therefore adjusted to the river network as described in Section 3.2.2. Second, WRF-Hydro-HAND relies on

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

predefined synthetic rating curves to convert routed discharge to stage. Uncertainties in these relationships may contribute to systematic bias, particularly where local channel geometry and roughness differ from assumptions embedded in the HAND dataset (Johnson et al., 2019). Finally, none of the models were calibrated in this study. This was an intentional choice to enable a more direct comparison of structural differences among the frameworks without the influence of model-specific parameter tuning. Consequently, the results represent uncalibrated model behavior rather than the best achievable performance of each modelling system, and relative model performance could differ following calibration.

6. Conclusion

This study demonstrates that substantial differences in simulated water level and flood extent arise from underlying model structural assumptions. Models that emphasize rainfall-driven overland flow, represented here by Torrent, tend to produce more spatially extensive inundation, whereas channel-based WRF-Hydro-HAND approache constrains flooding closer to the river network. HEC-RAS provides a more balanced representation of observed inundation for this event through its explicit treatment of rainfall-driven surface flow and channel-floodplain hydraulics. These differences are particularly pronounced in low-relief and floodplain environments, where storage and connectivity processes play a critical role.

By combining multi-model comparison with both point-based and spatial evaluation using satellite observations, this work highlights the importance of considering model structure when interpreting flood simulations. KGE′ decomposition reveals different contributions from bias, variability, and temporal agreement, while peak-timing and DSWE analyses provide complementary information on flood-wave propagation and spatial inundation. Together, these results emphasize the importance of considering model structure and multiple dimensions of performance when interpreting flood simulations. Future work should extend this framework to include additional modelling systems to better characterize structural variability across hydrologic and hydrodynamic models and to support flood risk assessment. The multi-model dataset developed in this study also provides a foundation for data-driven approaches, where machine learning models can leverage multiple physically based simulations to improve efficiency while retaining sensitivity to underlying processes.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Appendix A

Table A1 List of Station that used in this study. Coordinate Method describes adjustment method for location coordinate used in this study: (1) SWIM (2) nearest point at NHDPlus HR (3) original USGS coordinate (4) manual read from the map. Location shows in Figure A1. Order is the stream order in NHDPlus. Method for L describes the method for converting gage height to water level. Elevation is obtained from the DEM at the adjusted coordinates.

Site ID	Station Name	lon	lat	Coordinate Method	Offset from USGS (m)	Method for L	Drainage Area (km2)	Order	Elevation (m)	HUC 10
11461000	RUSSIAN R NR UKIAH	-123.19	39.2	SWIM	19.5	GZF	259	4	186	Headwaters Russian River
11461500	EF RUSSIAN R NRLPELLA	-123.13	39.25	HR snapped	13.5	GZF	238.8	4	244	East Fork Russian River
11462080	RUSSIAN R NR TALMAGE	-123.18	39.11	HR snapped	31.1	GZF	740.7	5	169	Upper Russian River
11462500	RUSSIAN R NR HOPLAND	-123.13	39.03	USGS	57.7	GZF	937.6	5	154	Upper Russian River
11463000	RUSSIAN R NR CLOVERDALE	-123.05	38.88	HR snapped	22.8	GZF	1302.8	5	116	Upper Russian River
11463170	BIG SULPHUR C A G RESORT NR CLOVERDALE	-122.8	38.8	SWIM	19	GZF	33.9	2	439	Big Sulphur Creek
11463200	BIG SULPHUR C NR CLOVERDALE	-123	38.83	SWIM	19.5	GZF	221.4	4	98	Big Sulphur Creek
11463500	RUSSIAN R A GEYSERVILLE	-122.9	38.71	USGS	18.5	GZF	1696.4	5	57	Middle Russian River
11463682	RUSSIAN R A JIMTOWN	-122.83	38.66	HR snapped	39.4	GZF	1771.6	5	45	Middle Russian River
11463900	MAACAMA C NR KELLOGG	-122.76	38.64	USGS		GZF	113.2	3	58	Middle Russian River
11463980	RUSSIAN R A DIGGER BEND NR HEALDSBURG	-122.86	38.63	HR snapped	40.8	GZF	2048.7	5	27	Middle Russian River
11464000	RUSSIAN R NR HEALDSBURG	-122.84	38.61	SWIM	22.3	GZF	2053.9	5	25	Middle Russian River
11465200	DRY C NR GEYSERVILLE	-122.96	38.7	SWIM	56.7	GZF	419.6	4	52	Dry Creek

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

11465240	DRY C BLW LAMBERT BR NR GEYSERVILLE	-122.93	38.65	HR snapped	21	GZF	453.2	4	37	Dry Creek
11465350	DRY C NR MOUTH NR HEALDSBURG	-122.86	38.59	USGS		GZF	562	4	19	Dry Creek
11465660	COPELAND C A ROHNERT PARK	-122.7	38.34	USGS	8.2	GZF	14.2	1	29	Mark West Creek
11465680	LAGUNA DE SANTA ROSA A STONY PT RD NR COTATI	-122.74	38.35	HR snapped	13.8	GZF	105.7	3	24	Mark West Creek
11465690	COLGAN C NR SANTA ROSA	-122.73	38.4	USGS		GZF	8.8	1	30	Mark West Creek
11465700	COLGAN C NR SEBASTOPOL	-122.77	38.37	HR snapped	17	GZF	17.6	1	22	Mark West Creek
11465750	LAGUNA DE SANTA ROSA C NR SEBASTOPOL	-122.83	38.43	HR snapped	25.4	DEM	206.2	3	18	Mark West Creek
11466170	MATANZAS C A SANTA ROSA	-122.7	38.44	HR snapped	5.4	GZF	54.4	3	54	Mark West Creek
11466200	SANTA ROSA C A SANTA ROSA	-122.72	38.44	HR snapped	23	GZF	147.6	4	41	Mark West Creek
11466320	SANTA ROSA C A WILLOWSIDE RD NR SANTA ROSA	-122.81	38.45	HR snapped	6.2	GZF	201	4	19	Mark West Creek
11466800	MARK WEST C NR MIRABEL HEIGHTS	-122.85	38.49	HR snapped	2.4	GZF	650.1	4	13	Mark West Creek
11467000	RUSSIAN R A HACIENDA BRIDGE NR GUERNEVILLE	-122.93	38.51	SWIM	30	GZF	3465.4	5	7	Lower Russian River
11467002	RUSSIAN R A JOHNSONS BEACH A GUERNEVILLE	-123	38.5	HR snapped	12.4	DEM	3504.3	5	4	Lower Russian River
11467200	AUSTIN C NRZADERO	-123.07	38.51	SWIM	16.8	GZF	162.7	4	14	Austin Creek
11467270	RUSSIAN R A HIGHWAY 1 BRIDGE NR JENNER	-123.1	38.43	Manual from map	106.1	DEM	3831.6	5	2	Lower Russian River

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Figure A1 Map of USGS gauge and HUC10 boundary

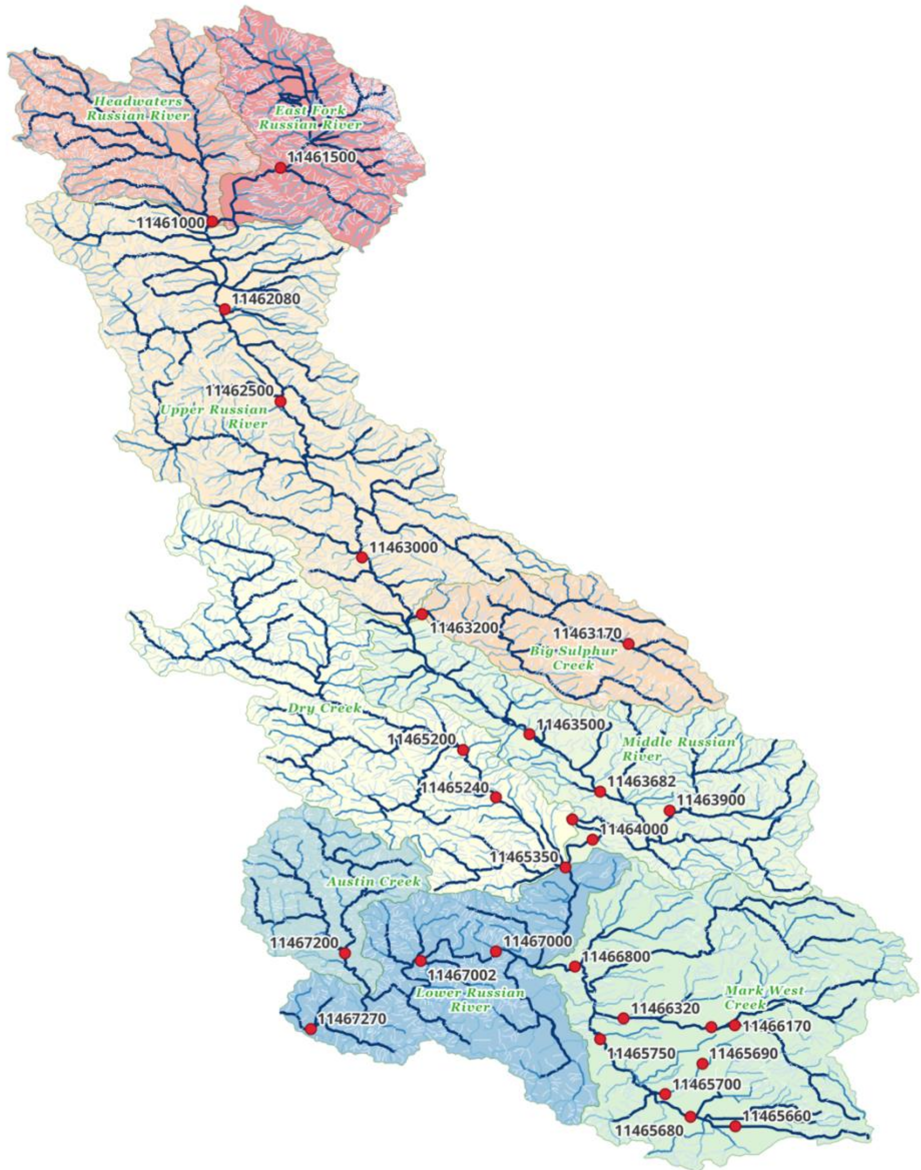
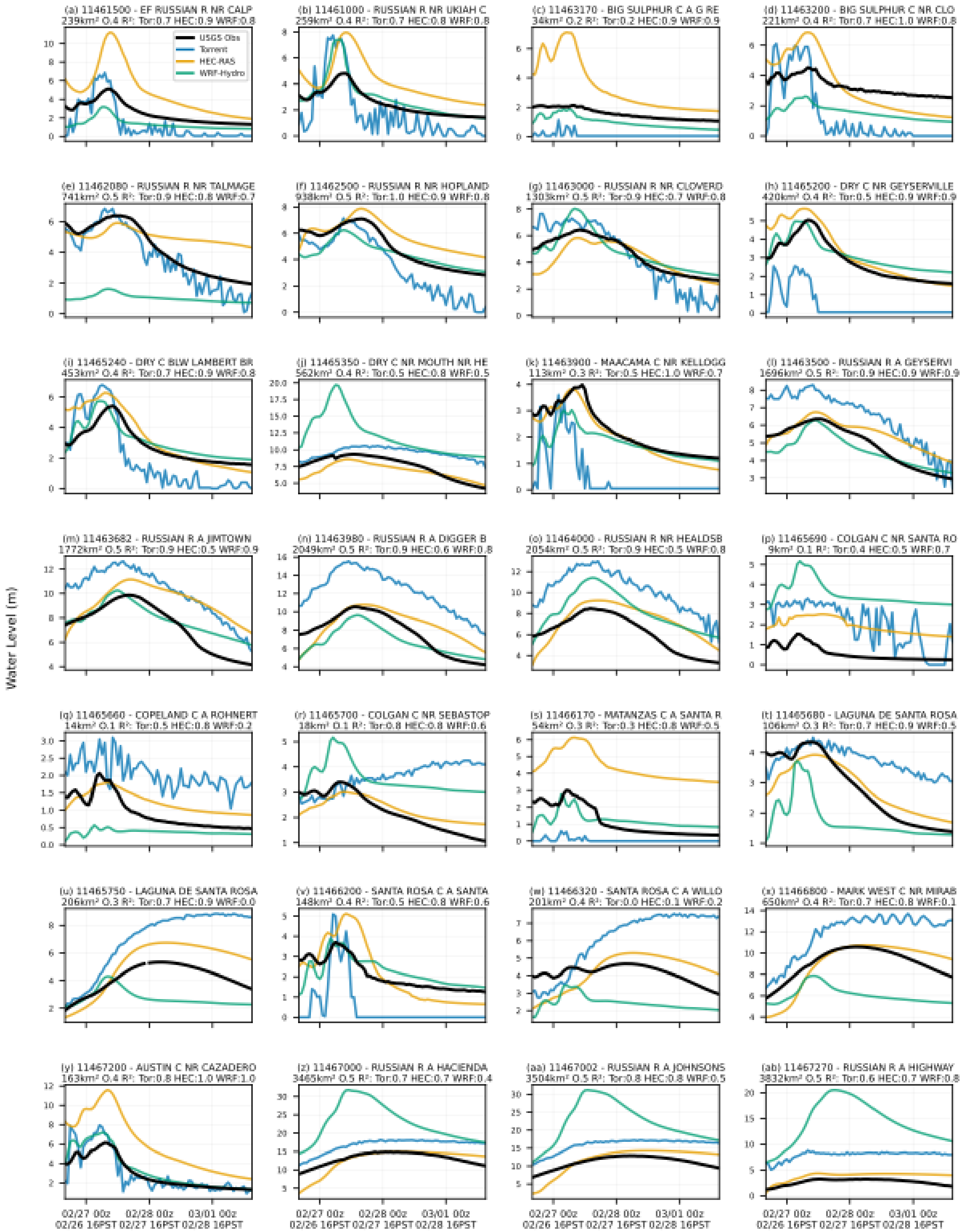


Figure A1. Locations of USGS stream gauges used in this study. Red markers indicate gauge locations, labelled by station ID. The study region corresponds to HUC8 18010110 and is subdivided into nine HUC10 basins, shown as colored polygons with HUC10 names labelled.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026



https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

545 **Figure A2. Timeseries of water level at all 28 gauges used in this study**

Appendix B

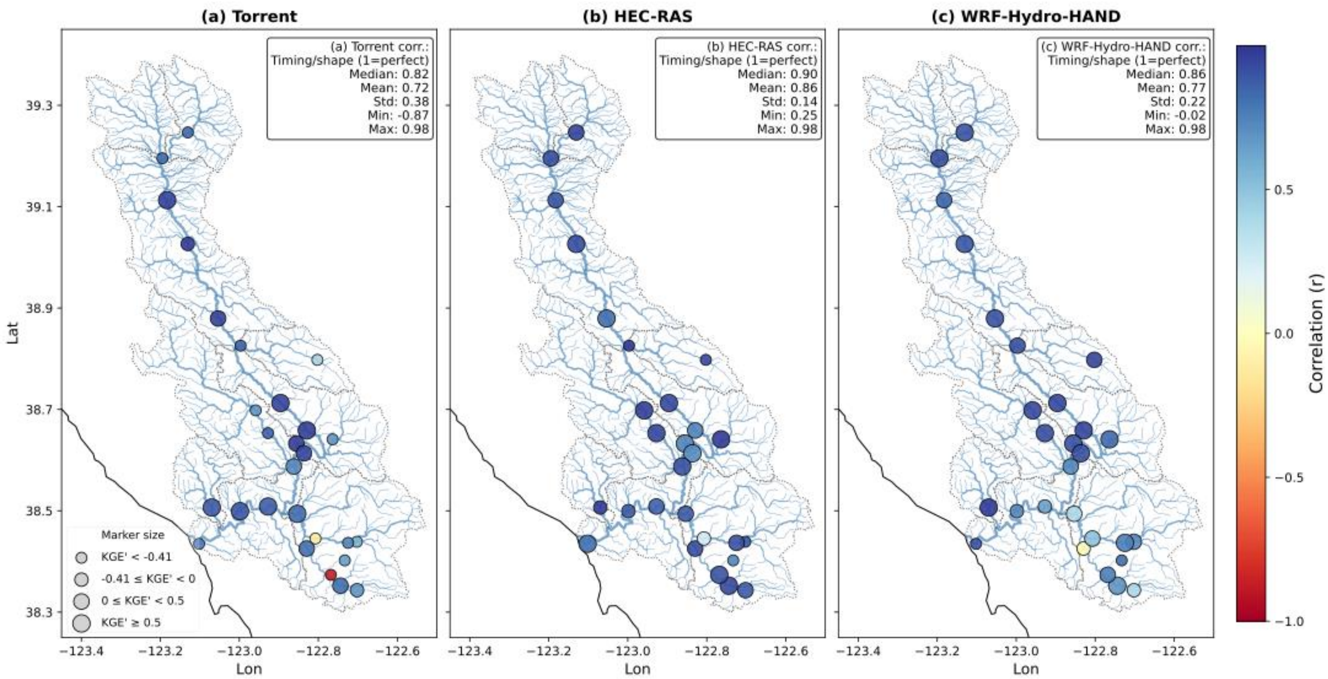
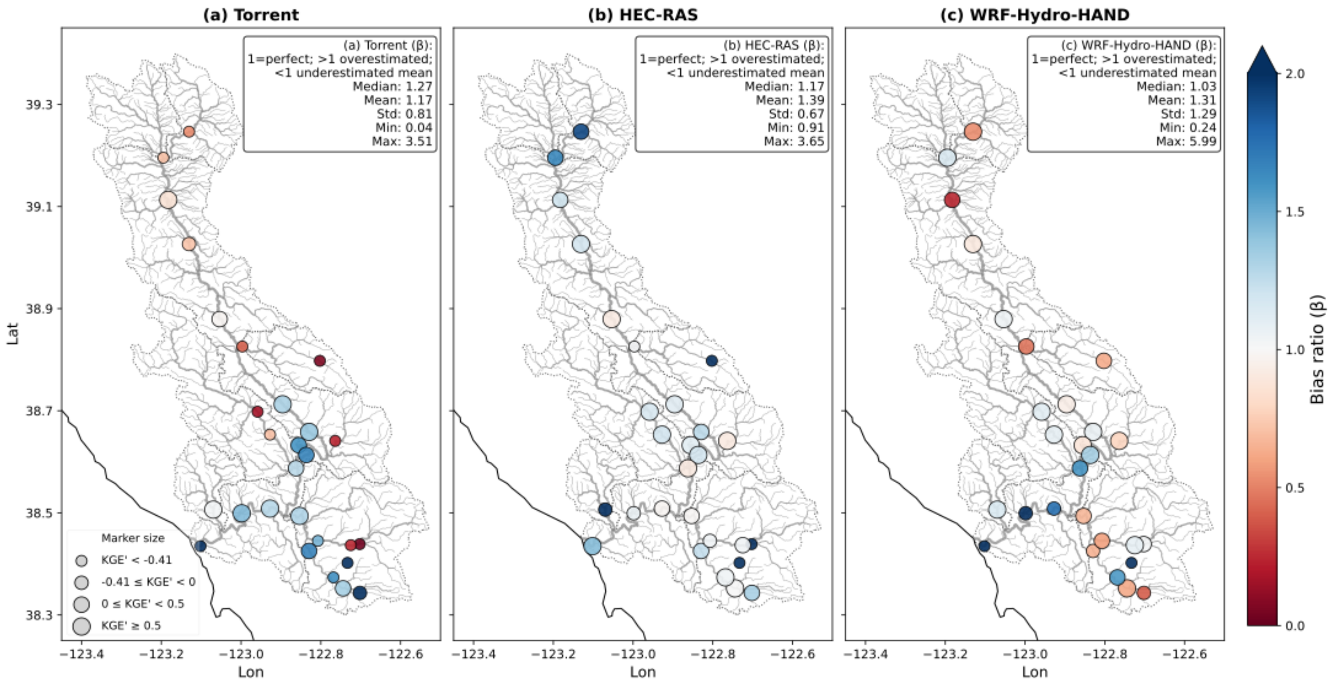


Figure B1. Correlation at gauges for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Marker color indicates correlation, and marker size is proportional to KGE′.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026



https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Figure B2. Bias ratio at gauges for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Marker color indicates bias ratio, and marker size is proportional to KGE′.

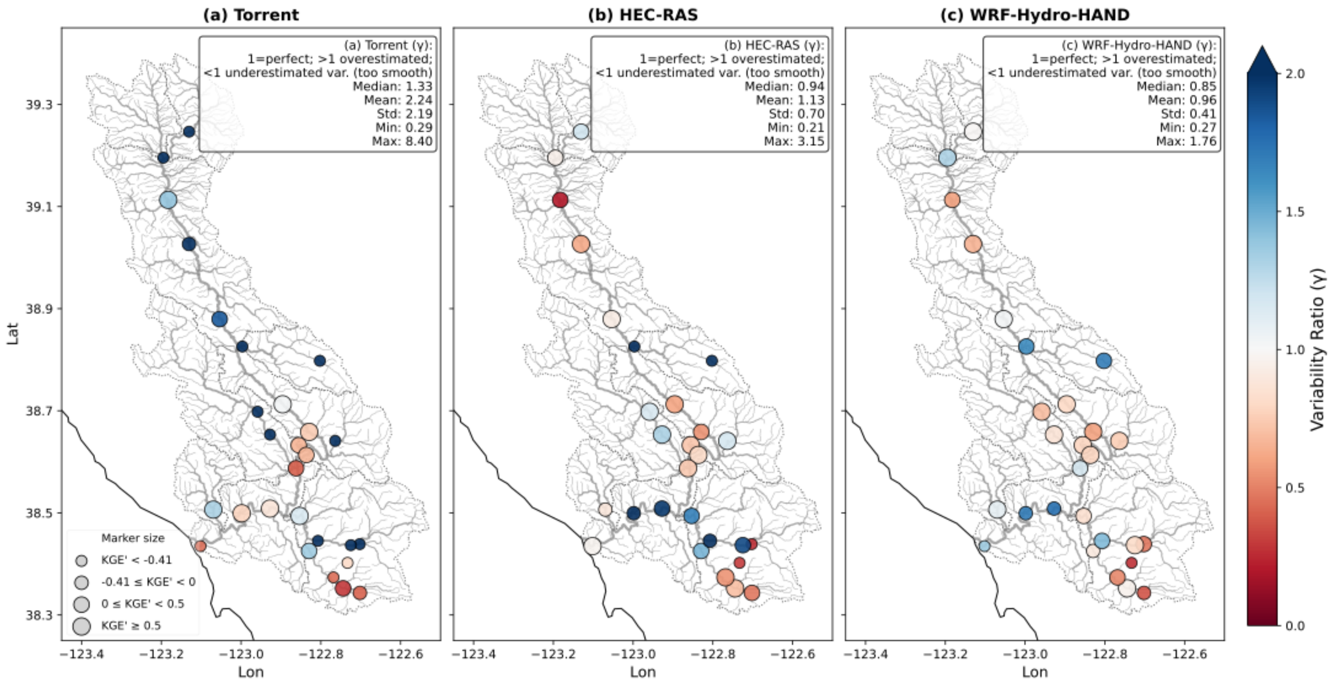


Figure B3. Variability ratio at gauges for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Marker color indicates variability ratio, and marker size is proportional to KGE′.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

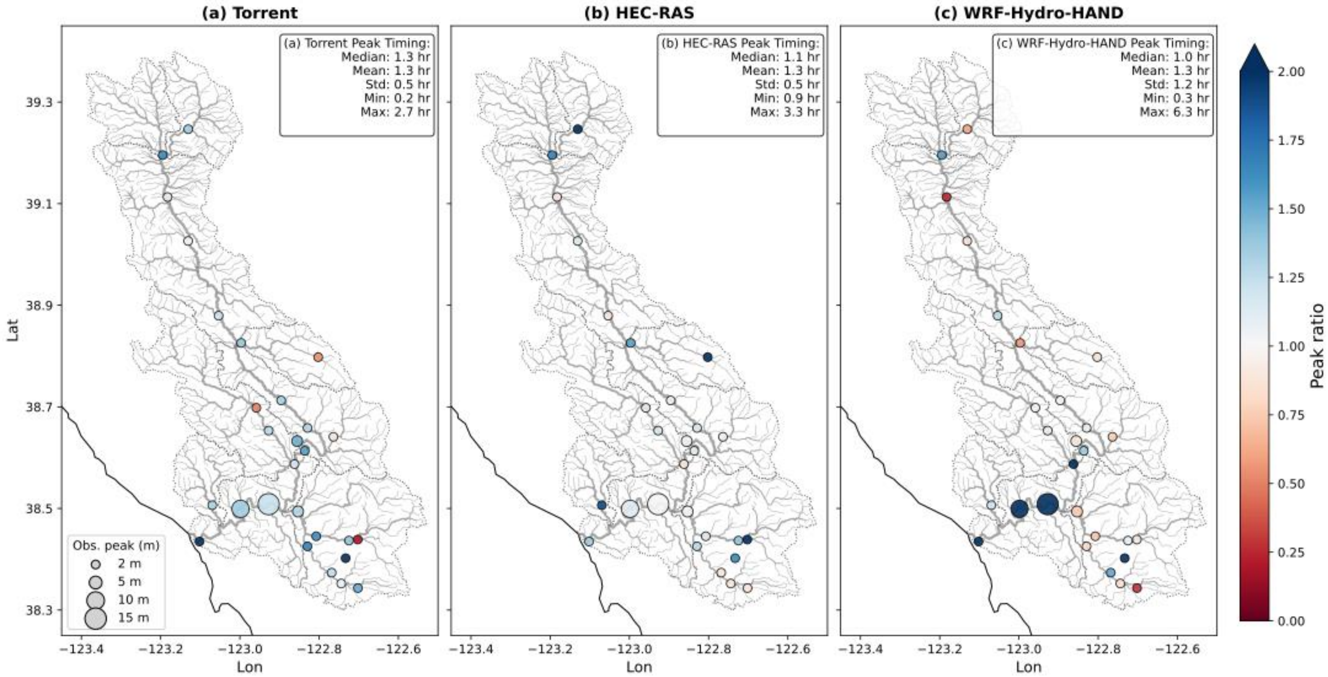


Figure B4. Peak bias ratio at gauges for (a) Torrent, (b) HEC-RAS (rain-on-grid), and (c) WRF-Hydro-HAND. Marker color indicates peak bias ratio, and marker size is proportional observed peak value.

Code and model availability

HEC-RAS 6.7 is publicly available from the U.S. Army Corps of Engineers (https://www.hec.usace.army.mil/software/hec-ras/). The Torrent model code is also publicly available at https://github.com/DOE-ICoM/Torrent.jl. WRF-Hydro model outputs used in this study were obtained from the publicly available NOAA National Water Model (NWM) retrospective dataset (https://registry.opendata.aws/nwm-archive/). Custom scripts used for data processing and analysis are available from the corresponding author upon reasonable request.

Author contributions

WYW and RL designed the study framework. WYW prepared the manuscript with contributions from all co-authors. SD and RL performed the Torrent simulations. MFM and RL conducted the HEC-RAS simulations. WYW processed the NWM reanalysis and HAND datasets and generated all figures. All authors contributed to the interpretation of results and the development of the conclusions.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

Competing interests

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

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-2019521. The authors acknowledge the NASA Disasters Program for providing satellite imagery. We thank the USGS for stream gauge observations via Dataretrieval python package and Landsat products via EarthExplorer, and the U.S. Army Corps of Engineers (USACE) for development of the HEC-RAS modelling system. We also acknowledge NOAA for providing the National Water Model (NWM) reanalysis dataset. The Torrent model was provided by its developers, and we thank Dr. W. Brent Daniel from PNNL for making the model available for this study. Artificial intelligence tools were used to assist with language polishing during manuscript preparation. All scientific interpretation, analysis, and conclusions were developed and verified by the authors.

Financial support

All authors have been supported by the LLNL-LDRD Program under Exploratory Research 25-ERD-035 “Flood Vulnerability”.

Reference

Afshari, S., Tavakoly, A. A., Rajib, M. A., Zheng, X., Follum, M. L., Omranian, E., & Fekete, B. M: Comparison of new generation low-complexity flood inundation mapping tools with a hydrodynamic model. Journal of Hydrology, 556, 539-556, 2018.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Brunner, G. W.: HEC-RAS River Analysis System: Hydraulic Reference Manual, Version 5.0, US Army Corps of Engineers–Hydrologic Engineering Center, 547, 2016.

Cifelli, R., Chandrasekar, V., Herdman, L., Turner, D. D., White, A. B., Alcott, T. I., Anderson, M., Barnard, P., Biswas, S. K., Boucher, M., Bytheway, J., Chen, H., Cutler, H., English, J. M., Erikson, L., Junyent, F., Gottas, D. J., Jasperse, J., Johnson, L. E., Krebs, J., Lindt, J. van de, Kim, J., Leon, M., Ma, Y., Marquis, M., Moninger, W., Pratt, G., Radhakrishnan, C., Shields, M., Spaulding, J., Tehranirad, B., and Webb, R.: Advanced Quantitative Precipitation Information: Improving Monitoring and Forecasts of Precipitation, Streamflow, and Coastal Flooding in the San Francisco Bay Area, https://doi.org/10.1175/BAMS-D-21-0121.1, 2024.

Clark, E. A., Sheffield, J., Vliet, M. T. H. van, Nijssen, B., and Lettenmaier, D. P.: Continental Runoff into the Oceans (1950–2008), Journal of Hydrometeorology, 16, 1502–1520, https://doi.org/10.1175/JHM-D-14-0183.1, 2015a.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, https://doi.org/10.1002/2015WR017198, 2015b.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, Journal of Hydrology, 375, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Congalton, R. G.: A review of assessing the accuracy of classifications of remotely sensed data, Remote Sensing of Environment, 37, 35–46, https://doi.org/10.1016/0034-4257(91)90048-B, 1991.

CW3E, CW3E AR Update: 25-27 February Post Event Summary, Center for Western Weather and Water Extremes, https://cw3e.ucsd.edu/cw3e-ar-update-25-27-february-post-event-summary/, 2026.

Daniel, B., Roth, C., Li, X., Rakowski, C., McPherson, T., and Judi, D.: Extremely rapid, Lagrangian modeling of 2D flooding: A rivulet-based approach, Environmental Modelling & Software, 161, 105630, https://doi.org/10.1016/j.envsoft.2023.105630, 2023.

Daniel, B.: Torrent.jl: Computationally efficient flood simulation using Lagrangian rivulets [Computer software]. U.S. Department of Energy. Retrieved from https://github.com/DOE-ICoM/Torrent.jl, 2025

Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., Trypaluk, C., Wu, W., and Miller, D.: The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation, JAWRA Journal of the American Water Resources Association, 59, 1246–1272, https://doi.org/10.1111/1752-1688.13143, 2023.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Gochis, D., Barlage, M., Cabell, R., Dugger, A., Fanfarillo, A., FitzGerald, K., McAllister, M., McCreight, J., RafieeiNasab, A., Read, L., Frazier, N., Johnson, D., Mattern, J. D., Karsten, L., Mills, T. J., and Fersch, B.: WRF-Hydro® v5.1.1, , https://doi.org/10.5281/ZENODO.3625238, 2020.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Han, H., Kim, J., Chandrasekar, V., Choi, J., and Lim, S.: Modeling Streamflow Enhanced by Precipitation from Atmospheric River Using the NOAA National Water Model: A Case Study of the Russian River Basin for February 2004, Atmosphere, 10, 466, https://doi.org/10.3390/atmos10080466, 2019.

Hayes, L., Chase, K. J., Wieczorek, M. E., and Jackson, S. E.: USGS streamgages in the conterminous United States indexed to NHDPlus v2.1 flowlines to support Streamgage Watershed InforMation (SWIM), 2021, U.S. Geological Survey data release [dataset], https://doi.org/10.5066/P9J5CK2Y, 2023.

Johnson, J. M., Munasinghe, D., Eyelade, D., and Cohen, S.: An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology, Natural Hazards and Earth System Sciences, 19, 2405–2420, https://doi.org/10.5194/nhess-19-2405-2019, 2019.

Johnson, J. M., Blodgett, D. L., Clarke, K. C., and Pollak, J.: Restructuring and serving web-accessible streamflow data from the NOAA National Water Model historic simulations, Sci Data, 10, 725, https://doi.org/10.1038/s41597-023-02316-7, 2023.

Jones, J. W.: Improved Automated Detection of Subpixel-Scale Inundation—Revised Dynamic Surface Water Extent (DSWE) Partial Surface Water Tests, Remote Sensing, 11, 374, https://doi.org/10.3390/rs11040374, 2019.

Kay, A. L., Davies, H. N., Bell, V. A., and Jones, R. G.: Comparison of uncertainty sources for climate change impacts: flood frequency in England, Climatic Change, 92, 41–63, https://doi.org/10.1007/s10584-008-9471-4, 2009.

Kim, J., Cho, E., Chandrasekar, V., Cifelli, R., Johnson, L. E., Wood, A., and Gochis, D.: A regional assessment of the operational National Water Model in the San Francisco Bay area for the 2018–2019 water year, Journal of Hydrology, 660, 133362, https://doi.org/10.1016/j.jhydrol.2025.133362, 2025.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Knoben, W. J., Freer, J. E., & Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23(10), 4323-4331, 2019

Liu, Y., Tarboton, D., and Maidment, D.: Height Above Nearest Drainage (HAND) and Hydraulic Property Table for CONUS, https://doi.org/10.13139/ORNLNCCS/1608331, 2020.

Liu, Y. Y., Maidment, D. R., Tarboton, D. G., Zheng, X., and Wang, S.: A CyberGIS Integration and Computation Framework for High-Resolution Continental-Scale Flood Inundation Mapping, JAWRA Journal of the American Water Resources Association, 54, 770–784, https://doi.org/10.1111/1752-1688.12660, 2018.

Ma, Y., Chandrasekar, V., Chen, H., and Cifelli, R.: Quantifying the Potential of AQPI Gap-Filling Radar Network for Streamflow Simulation through a WRF-Hydro Experiment, https://doi.org/10.1175/JHM-D-20-0122.1, 2021.

NASA: Flooding on the Russian River, NASA Earth Observatory, available at: https://science.nasa.gov/earth/earth-observatory/flooding-on-the-russian-river-144619/ (last access: 10 August 2026), 2019.

Nikrou, P., Seyvani, S., Cohen, S., Gangrade, S., Gutenson, J.L., Tian, D., Baruah, A., Follum, M.L. and Kao, S.C.: Intercomparison of flood inundation models across land use types and hydrological flood stages, Journal of Hydrology, p.135410, 2026

NOAA NCEI: U.S. Billion-dollar Weather and Climate Disasters, 1980 - present (NCEI Accession 0209268): https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:0209268, last access: 3 May 2024.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, https://doi.org/10.1029/2009WR008328, 2010.

Thirel, G., Andréassian, V., and Perrin, C.: On the need to test hydrological models under changing conditions, Hydrological Sciences Journal, 60, 1165–1173, https://doi.org/10.1080/02626667.2015.1050027, 2015.

https://doi.org/10.5194/egusphere-2026-3064
Preprint. Discussion started: 31 August 2026

U.S. Army Corps of Engineers (USACE), 2026. HEC-RAS Hydraulic Reference Manual, Version 6.7. Hydrologic Engineering Center, Davis, CA. https://www.hec.usace.army.mil/confluence/rasdocs/r2dum/latest, last access: 13 April 2026.

U.S. Geological Survey, National Water Information System data available on the World Wide Web (USGS Water Data for the Nation) [dataset], https://waterdata.usgs.gov/nwis/, 2020a

U.S. Geological Survey, Earth Resources Observation and Science (EROS) Center Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2 [dataset], https://doi.org/10.5066/P9OGBGM6, 2020b

U.S. Geological Survey, USGS National Hydrography Dataset Plus High Resolution National Release 1 [dataset], https://doi.org/10.5066/P9WFOBQI, 2022

U.S. Geological Survey, Annual NLCD Collection 1 Science Products: U.S. Geological Survey data release [dataset], https://doi.org/10.5066/P94UXNTS, 2024

U.S. Geological Survey, 3D Elevation Program (3DEP), 1/3 arc-second digital elevation model (DEM) [website], https://www.usgs.gov/3d-elevation-program, last access: 13 April 2025.

Wright, D. B., Bosma, C. D., and Lopez-Cantu, T.: U.S. Hydrologic Design Standards Insufficient Due to Large Increases in Frequency of Rainfall Extremes, Geophysical Research Letters, 46, 8144–8153, https://doi.org/10.1029/2019GL083235, 2019.

Ye, A., Zhou, Z., You, J., Ma, F., & Duan, Q.: Dynamic Manning's roughness coefficients for hydrological modelling in basins, Hydrology Research, 49(5), 1379-1395, 2018.

Zheng, X., Maidment, D. R., Tarboton, D. G., Liu, Y. Y., and Passalacqua, P.: GeoFlood: Large-Scale Flood Inundation Mapping Based on High-Resolution Terrain Analysis, Water Resources Research, 54, 10,013-10,033, https://doi.org/10.1029/2018WR023457, 2018.