



1 **TORRENT: An observation-constrained and benchmark-ready community shared**
2 **dataset in the contiguous United States for flash-flood research and model**
3 **intercomparison**

4 Songkun Yan¹, Mengye Chen¹, Jonathan J. Gourley², Jian Zhang², Günter Blöschl³, Paul Bates⁴, Zhi
5 Li⁵, Siyu Zhu⁶, Abigail Snooks¹, Yixin Wen⁷, Xiaodong Chen^{1,8}, Chengbin Deng^{9,10}, Juan Camilo
6 Parra Toro¹¹, Meihong Ma¹², Yang Hong¹

7 ¹School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK,
8 73072, USA

9 ²NOAA/OAR National Severe Storms Laboratory, Norman, OK, 73072, USA

10 ³Institute of Hydraulic and Water Resources Engineering, Vienna University of Technology, Vienna,
11 1040, Austria

12 ⁴School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK

13 ⁵Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, Boulder, CO,
14 80304, USA

15 ⁶Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

16 ⁷Department of Geography, University of Florida, Gainesville, FL, USA

17 ⁸School of Meteorology, University of Oklahoma, Norman, OK, USA

18 ⁹Center for Spatial Analysis, University of Oklahoma, Norman, OK, USA

19 ¹⁰Department of Geography and Environmental Sustainability, University of Oklahoma, Norman,
20 OK, USA

21 ¹¹Research Group GIS, Politécnico Colombiano Jaime Isaza Cadavid, Facultad de Ingeniería,
22 Medellín 050022, Colombia

23 ¹²Faculty of Geography, Tianjin Normal University, Tianjin 300387, China

24 Corresponding Author: Yang Hong Email: yanghong@ou.edu; Meihong Ma Email:
25 mamh@tjnu.edu.cn

26



27 Abstract

28 Flash floods develop within minutes to hours, occur mainly in small to medium-sized watersheds,
29 and are documented through scattered sources such as news reports, social media, stream gauges,
30 operational storm events databases, and post-flood field surveys. Such fragmentation historically
31 limits flash flood research advances and model evaluation or intercomparisons. We present
32 **TORRENT** (the **T**raceable, **O**bservation-constrained, **R**apid-Response, **E**pisode-gauge-watershed
33 **N**etwork of **T**estbeds), a benchmark-ready community dataset of flash-flood hydrological response
34 in the contiguous United States (CONUS). Each testbed links one reported NOAA/NCEI flash-flood
35 episode, one U.S. Geological Survey (USGS) stream gauge, and the gauge's contributing watershed,
36 combining the observed discharge hydrograph—the primary benchmark target—with watershed
37 boundaries, temporally screened in-basin high-water marks, and quality-control metadata; full
38 packages additionally provide Multi-Radar Multi-Sensor (MRMS) precipitation forcing and
39 auxiliary gridded products. TORRENT follows a three-level structure distinguishing what was
40 reported, what observations are available, and what can be evaluated reproducibly: a national
41 inventory of 42,466 reported episodes for 1996–2025; observation-availability documentation for
42 5,424 recent CONUS episodes during 2021–2025, including episodes without usable external
43 observations; and 806 episode-gauge-watershed evaluation packages representing 160 episodes
44 and 208 gauges. Each Level-3 package has a retrievable watershed, at least one in-basin high-water
45 mark meeting the stated temporal-screening criterion, and a detected discharge response. Flash-
46 flood-response screening, deduplication, in-basin NCEI corroboration, station-relative severity
47 screening, and post-selection curation yield a conservative primary benchmark of 21 testbeds
48 representing 14 episodes and 21 gauges. This primary set is intended as a high-confidence reference
49 for formal rainfall-runoff model intercomparison rather than as a statistically representative sample
50 of CONUS flash floods; the broader Level-3 pool supports exploratory analysis and robustness
51 testing. Primary testbeds exhibit rapid, pronounced responses, with estimated event-peak return
52 periods ranging from about 2 to more than 100 years. The automated construction and screening
53 stages are released as deterministic, re-runnable code, together with documented curation decisions,
54 sensitivity tests, and quality-control diagnostics. TORRENT is openly available at
55 <https://doi.org/10.5281/zenodo.23359069> (Yan et al., 2026).

56



57 1 Introduction

58 Flash floods are among the most difficult hydrological hazards to observe, reconstruct, and predict.
 59 They often develop within minutes to a few hours after intense rainfall, frequently occur in small to
 60 medium-sized watersheds, and are controlled by nonlinear interactions among rainfall intensity,
 61 storm motion, terrain, land cover, channel routing, soil conditions, and antecedent wetness (Borga
 62 et al., 2011; Gourley et al., 2017; Hapuarachchi et al., 2011; Marchi et al., 2010). Their hydrological
 63 signatures are highly localized: a severe flash flood may be visible at one stream gauge, absent at
 64 another nearby gauge, reported only through county-level storm documentation, or physically
 65 recorded only through post-event field evidence such as high-water marks. This distinguishes many
 66 flash-flood cases from larger river floods, for which broader flood waves and longer response times
 67 often facilitate event attribution and spatial aggregation.

68 Despite major progress in radar precipitation estimation (Zhang et al., 2016), distributed
 69 hydrological modeling (Li et al., 2023a), and operational flash-flood monitoring (Gourley et al.,
 70 2017), systematic flash-flood response studies remain constrained by fragmented evidence. National
 71 storm-event archives provide broad and long-term records of reported impacts, but reported
 72 locations may be spatially imprecise, administratively defined, or distributed across multiple
 73 counties and forecast zones. Stream-gauge networks provide direct hydrological observations, but
 74 a hydrograph peak alone does not identify the responsible storm or prove that reported flash-flood
 75 impacts occurred within the contributing watershed. Radar-based precipitation products provide
 76 spatially continuous forcing, but their interpretation depends on the correct event window,
 77 watershed boundary, outlet location, and response target (Gourley et al., 2017; Zhang et al., 2016).
 78 Post-flood high-water marks provide independent physical evidence of inundation (Koenig et al.,
 79 2016), but they are collected and archived separately from storm reports, streamflow time series,
 80 and gridded precipitation fields. As a result, many flash-flood studies still require case-by-case
 81 assembly and manual interpretation, limiting reproducibility and making it difficult to compare
 82 events, regions, watersheds, and models.

83 This fragmentation is especially problematic for model evaluation and intercomparison. Event-scale
 84 rainfall–runoff models and data-driven hydrological prediction systems increasingly rely on
 85 benchmark datasets that pair forcing, catchment attributes, and observed streamflow response in
 86 reproducible formats. Large-sample hydrology datasets such as MOPEX, CAMELS, CAMELS-GB,
 87 CAMELS-BR, CAMELS-AUS, LamaH-CE, Caravan, and related continental-scale collections
 88 have transformed rainfall–runoff modeling by enabling standardized comparison across many
 89 watersheds (Addor et al., 2017; Arsenault et al., 2020; Chagas et al., 2020; Coxon et al., 2020; Duan
 90 et al., 2006; Fowler et al., 2021; Klingler et al., 2021; Kratzert et al., 2023; Newman et al., 2015).
 91 However, most large-sample hydrology datasets are organized around continuous daily or hourly
 92 catchment time series for long-term water resources management, not event-scale flash-flood
 93 episodes constrained by storm reports, high-water marks, radar precipitation, and gauge-level flood
 94 response. Consequently, there remains a need for a dataset that is not merely a list of reported flash
 95 floods and not merely a streamflow archive, but a traceable collection of episode–gauge–watershed
 96 units suitable for mechanistic understanding of flash-flood response and model intercomparison.
 97 The principal use considered here is event-scale rainfall–runoff response evaluation; hydraulic
 98 inundation modelling and land-surface-model evaluation require additional observations and
 99 constraints beyond the core benchmark package.



100 The need for such a resource is increasing as short-duration rainfall extremes, flood exposure, and
 101 observation availability change across regions (Bates, 2022; Li et al., 2022; Zhu et al., 2026).
 102 Because observed flood occurrence and impacts reflect both physical processes and uneven
 103 monitoring infrastructure, an event-scale benchmark must preserve the connection among
 104 precipitation forcing, watershed response, and independent flood evidence rather than assume that
 105 reported events form a spatially representative sample (Merz et al., 2021; Slater et al., 2021; Tellman
 106 et al., 2021; Wing et al., 2022). For flash floods, this requires documenting whether the relevant
 107 observations exist and whether they are spatially and temporally consistent with the selected
 108 watershed and gauge response.

109 Here we present TORRENT (the Traceable, Observation-constrained, Rapid-Response, Episode-
 110 gauge-watershed Network of Testbeds), a benchmark-ready community data resource for flash-
 111 flood hydrological response in the contiguous United States. In this study, a testbed refers to a static,
 112 self-contained episode-gauge-watershed data package assembled for reproducible hydrological
 113 model evaluation, rather than an operational testing environment. An NCEI episode may comprise
 114 multiple flash-flood event records, each potentially associated with multiple reported locations. At
 115 L3, each testbed links one such episode to one USGS gauge and its contributing watershed;
 116 consequently, a single episode may produce multiple testbeds. TORRENT integrates NOAA/NCEI
 117 storm reports, USGS stream-gauge observations, USGS/FEMA Short-Term Network high-water
 118 marks, USGS NLDI watershed boundaries, USGS 3DEP terrain data, NOAA/NSSL MRMS
 119 precipitation forcing, and auxiliary model-derived CREST MaxUnitStreamflow fields. Its three-
 120 level structure distinguishes what was reported, what observations are available, and what can be
 121 evaluated reproducibly. Level 1 (L1) preserves the national reported-event inventory; Level 2 (L2)
 122 documents observation availability for recent CONUS episodes without treating observation gaps
 123 as exclusions; and Level 3 (L3) links reported episodes, gauges, contributing watersheds, discharge
 124 response, and independent flood evidence into reproducible evaluation packages, with gridded
 125 precipitation forcing included in full packages. A conservative primary benchmark of 21 testbeds
 126 representing 14 episodes and 21 gauges is intended as a high-confidence reference for formal
 127 rainfall-runoff model intercomparison, while the broader Level-3 pool supports exploratory
 128 analysis and robustness testing. The dataset is observation-constrained and is not intended as a
 129 statistically representative climatology of CONUS flash floods. The following sections describe the
 130 source datasets, construction and screening workflow, benchmark characteristics, quality-control
 131 diagnostics, recommended uses, and an example model application.

132 **2 Source datasets**

133 The benchmark data resource was constructed by integrating public storm-event records, stream-
 134 gauge observations, field-surveyed high-water marks, watershed and terrain datasets, gridded
 135 precipitation forcing, and auxiliary model-derived flood-response products. These source datasets
 136 serve complementary roles in the workflow: the NOAA National Centers for Environmental
 137 Information (NCEI) Storm Events Database defines the reported flash-flood episode population; the
 138 U.S. Geological Survey (USGS) National Water Information System (NWIS) provides the observed
 139 hydrological response and the annual peak-flow records used for flood-frequency reference;
 140 USGS/FEMA Short-Term Network (STN) high-water marks provide independent physical flood
 141 evidence; the USGS Network-Linked Data Index (NLDI) defines the gauge-linked watershed unit;
 142 the NOAA/NSSL Multi-Radar Multi-Sensor (MRMS) system provides precipitation forcing; USGS



143 3DEP provides terrain context; and the NOAA/NSSL FLASH CREST MaxUnitStreamflow product
144 provides auxiliary distributed-response context. Wildland Fire Interagency Geospatial Services
145 (WFIGS) fire perimeters are additionally used for post-selection assessment of recently burned
146 watersheds.

147 All source datasets are publicly accessible through agency data services and public archives, and
148 Table 1 summarizes the formats, native resolutions, and roles of the data used in the workflow. The
149 L1 reported-event inventory spans 1996–2025, whereas the L2 observation-availability catalogue
150 and L3 evaluation packages are restricted to 2021–2025 to provide a recent, internally consistent,
151 and practically retrievable observation period. This restriction reduces complications associated
152 with changes in radar products, stream-gauge metadata, field-observation availability, and quality-
153 control burden. The recent-period inventory contains 5,424 L2 episodes and ultimately supports 806
154 L3 evaluation packages; subsequent screening and post-selection curation define the primary
155 benchmark of 21 testbeds representing 14 episodes and 21 gauges. Detailed construction and
156 screening criteria are described in Sect. 3 and Appendix Table A1.



Table 1. Source datasets used to construct the TORRENT data resource. Storm-event records define the reported flash-flood episode population; stream-gauge, peak-flow, and high-water-mark data provide observational constraints; watershed and terrain data define the hydrological unit; gridded precipitation provides meteorological forcing; and auxiliary layers (CREST MaxUnitStreamflow and WFIGS fire perimeters) support spatial interpretation and post-selection curation.

Source dataset	Provider	Format	Native scale / resolution	Main role in workflow
Storm Events Database — event detail records	NOAA National Centers for Environmental Information	CSV	Event-level, 1996–2025	Provides the national flash-flood episode catalogue, including event type, episode ID, timing, location metadata, casualties, damages, causes, and narratives.
Storm Events Database — location records	NOAA National Centers for Environmental Information	CSV	Reported event-location points	Provides reported event-location points used to refine episode spatial footprints and support spatial matching with gauges, high-water marks, and watersheds.
NWIS instantaneous-value streamflow and stage	U.S. Geological Survey National Water Information System	JSON	Native sub-hourly IV records; sampling interval varies by gauge	Provides observed hydrological response at stream gauges for screening, primary-gauge selection, hydrograph validation, and benchmark target time series.
NWIS annual peak-flow records	U.S. Geological Survey National Water Information System	RDB text	Annual peak series by gauge	Provides station-relative flood-frequency context for LP3 return-period and upper-tail diagnostics.
Short-Term Network high-water marks	U.S. Geological Survey / Federal Emergency Management Agency	JSON	Field-surveyed point observations	Provides field-surveyed flood evidence used for observation-support assessment, watershed attribution, shared-event identification, and optional spatial diagnostics.
Network-Linked Data Index (NLDI) watershed boundaries	U.S. Geological Survey	GeoJSON	Gauge-linked watershed polygons	Provides gauge-linked watershed polygons that define the hydrological unit for precipitation extraction, terrain clipping, and gauge–event interpretation.
Three-dimensional Elevation Program (3DEP) digital elevation model	U.S. Geological Survey National Map	GeoTIFF	~10 m (1/3 arc-second) terrain grid	Provides terrain data for each watershed package and supports diagnostic mapping and terrain-based hydrological analysis.
Multi-Radar Multi-Sensor (MRMS) PrecipRate	NOAA National Severe Storms Laboratory, accessed through Iowa State University MT Archive	GRIB2	~1 km, 2 min	Provides high-resolution precipitation forcing for each benchmark watershed and event window, including gridded precipitation and basin-mean precipitation time series.
CREST MaxUnitStreamflow	NOAA National Severe Storms Laboratory FLASH system	GRIB2	~1 km, 10 min	Provides an auxiliary distributed flood-response layer for visualization and exploratory spatial interpretation; it is not treated as an independent observation or benchmark target.
WFIGS Interagency Fire Perimeters	Wildland Fire Interagency Geospatial Services, hosted by the National Interagency Fire Center	GeoJSON	Incident-perimeter polygons	Supports recent-wildfire applicability review and post-fire attribution metadata.



162 2.1 Storm-event records

163 The NOAA/NCEI Storm Events Database provides the source population of reported flash-flood
 164 episodes used in this study. The database contains event records compiled from National Weather
 165 Service reports and includes event type, episode identifier, begin and end times, state and county,
 166 Weather Forecast Office, reported casualties, property and crop damage estimates, flood causes, and
 167 narrative descriptions. Annual Storm Events detail files for 1996–2025 were used, and records with
 168 EVENT_TYPE equal to “Flash Flood” were retained.

169 The event detail records were complemented by the corresponding Storm Events location files,
 170 which provide reported event-location points associated with individual storm reports. These
 171 location records were used to refine the spatial footprint of each episode and to support spatial
 172 matching with stream gauges, high-water marks, and watershed boundaries. When location records
 173 were unavailable or incomplete, event-level begin and end coordinates were used as fallback spatial
 174 information. Together, the detail and location records define the national reported flash-flood
 175 episode catalogue, or Level 1 catalogue, and supply the event metadata used throughout the
 176 downstream screening workflow. The data were obtained from the NOAA/NCEI bulk Storm Events
 177 CSV archive (<https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/>).

178 2.2 USGS stream-gauge observations

179 Observed hydrological response was obtained from the U.S. Geological Survey National Water
 180 Information System (NWIS). Instantaneous-value discharge and stage observations were retrieved
 181 for candidate stream gauges using NWIS web services. Discharge records use parameter code 00060,
 182 and stage records use parameter code 00065. For each candidate episode–gauge pair, the retrieval
 183 window extended from 24 h before the reported episode begin time to 72 h after the reported episode
 184 end time, allowing the pre-event baseline, event peak, and post-peak recession to be evaluated. The
 185 window is a practical analysis choice for characterizing antecedent conditions and the post-peak
 186 recession, not a separate definition of flash-flood duration.

187 NWIS observations were accessed through the USGS Water Services instantaneous-values and site-
 188 metadata services. Stage observations were used for observation-availability assessment and
 189 supporting context where available, whereas the discharge hydrograph is the primary benchmark
 190 response target. Discharge records were used to determine whether a discernible flood response
 191 occurred and to calculate baseline discharge, peak discharge, peak-to-baseline ratio, response timing,
 192 and flashiness diagnostics. The public NWIS Water Services entry point is
 193 <https://waterservices.usgs.gov/>.

194 Annual instantaneous-peak records were retrieved separately from NWIS to provide station-relative
 195 flood-frequency context. These records were used to estimate the log-Pearson type III return period
 196 and upper-tail diagnostics for each event peak. The resulting quantities are treated as severity and
 197 quality-control descriptors rather than as independent benchmark targets. Extreme-event records
 198 may include estimated or indirectly determined values when direct measurements are unavailable
 199 or gauges are overtopped or damaged. Original NWIS qualifiers are preserved, and affected testbeds
 200 are identified through benchmark quality-control flags.

201 2.3 USGS/FEMA Short-Term Network high-water marks



High-water marks (HWMs) from the USGS/FEMA Short-Term Network (STN) provide independent physical evidence of flood occurrence and peak water level. STN HWMs are field-surveyed point observations that include geographic coordinates, surveyed elevation, vertical-datum metadata, quality grades, and event associations. They are not continuous hydrographs or complete inundation-extent observations; instead, they provide field evidence used to constrain event attribution and support optional spatial-response diagnostics. A high-water mark establishes flood occurrence but does not, by itself, distinguish a rapid-onset flash flood from a slower riverine flood.

HWM data were accessed through the USGS STN filtered HWM service (<https://stn.wim.usgs.gov/>). During L2 annotation, nearby HWMs were assigned temporal-screening classes using a window from 7 d before the NCEI episode began to 30 d after it ended. A mark was labeled “confirmed” when its parent STN event interval overlapped this window (using the event start plus 1 d if its end was missing), or when its survey date fell within the window if STN event dates were absent. Dated nonmatches were labeled “inconsistent”, and records with neither date were “uncertain”. These classes indicate possible survey-level temporal association, not verification of the individual gauge flood peak. During L3 qualification, a candidate episode–gauge pair was required to contain at least one in-basin HWM labeled “confirmed” by this screening rule. Thus, final qualification depends on watershed containment rather than proximity alone.

Shared HWM surveys and the number of confirmed in-basin HWMs provide supporting evidence during duplicate attribution and deterministic representative selection. Duplicate testbeds are ultimately identified using the gauge and flood-peak time. Original HWM elevations, quality grades, event associations, and vertical-datum fields are preserved, and records with missing or uncertain datum information are explicitly flagged.

2.4 USGS NLDI watershed boundaries and river-network context

Watershed boundaries were obtained using the USGS Network-Linked Data Index (NLDI). For each candidate stream gauge, the NLDI basin service provides the complete upstream contributing-watershed polygon associated with the registered gauge location. These polygons define the hydrological analysis unit for each episode–gauge–watershed testbed.

The NLDI watershed is used to clip and summarize gridded precipitation and auxiliary CREST data, calculate basin-mean precipitation time series, and determine whether HWM and NCEI report locations fall inside the gauge’s contributing watershed. It also provides the spatial frame for diagnostic mapping and gauge–event interpretation. Proximity to an episode report is used only to generate candidate gauges; final hydrological relevance is determined from the delineated watershed and in-basin evidence. Watershed boundaries were retrieved from the USGS NLDI basin service: <https://api.water.usgs.gov/nldi/linked-data/nwissite/USGS-{STADID}/basin>.

2.5 USGS 3DEP digital elevation model

Terrain data were obtained from the USGS 3D Elevation Program (3DEP). For each full L3 testbed, the approximately 10 m, 1/3-arc-second elevation layer was exported for the watershed domain and stored as a GeoTIFF. Over the CONUS, elevations are referenced to the North American Vertical Datum of 1988 (NAVD88), providing consistency with the majority of surveyed STN HWM elevations.



242 The DEM is used primarily for terrain context, diagnostic mapping, and interpretation of watershed
 243 relief, channel setting, and drainage structure. It also supports spatial comparison among the
 244 watershed boundary, gauge location, NCEI reports, and HWMs. Local elevation uncertainty varies
 245 with terrain and source-data quality, and the DEM does not explicitly represent channel bathymetry
 246 or hydraulic structures. It is therefore not treated as a complete terrain input for hydrodynamic
 247 inundation modeling. The 3DEP data were accessed through The National Map 3DEP ImageServer:
 248 <https://elevation.nationalmap.gov/arcgis/rest/services/3DEPElevation/ImageServer>.

249 **2.6 MRMS precipitation forcing**

250 Meteorological forcing was derived from NOAA/NSSL Multi-Radar Multi-Sensor (MRMS)
 251 products (Zhang et al., 2016). The MRMS data provide high-resolution gridded precipitation
 252 estimates over the event window and are used as the primary precipitation forcing in full L3
 253 packages. The workflow retrieves archived MRMS precipitation fields, extracts a window defined
 254 by the NLDI watershed bounding box without additional padding, and stores the resulting grid as a
 255 Zarr array. The forcing window extends from 24 h before episode onset to 6 h after its end. The
 256 source grid values are extracted without interpolation or aggregation; the released Zarr coordinates
 257 are reconstructed from the bounding box and may differ by up to approximately half a grid cell from
 258 source-grid centers.

259 For each full gauge testbed, MRMS data are used to compute basin-mean precipitation time series
 260 and event-total accumulated precipitation fields. The basin-mean time series supports hydrograph
 261 timing analysis, including rainfall-peak to streamflow-peak lag, while the event-total field provides
 262 spatial context for the storm forcing over the watershed. The MRMS PrecipRate_00.00 fields were
 263 retrieved from the Iowa State MT Archive (<https://mtarchive.geol.iastate.edu/>) as archived
 264 operational NCEP products. Software-version identifiers were not retained in the input files, and no
 265 retrospective harmonization across operational versions was performed. Radar-derived precipitation
 266 estimates can be affected by radar range, beam blockage and overshooting, precipitation
 267 microphysics, and terrain-related sampling limitations, particularly during extreme rainfall and in
 268 complex terrain. Hourly MRMS Radar Quality Index (RQI) layers provide complementary radar-
 269 sampling quality information (Sect. 5.2), while event-level precipitation–runoff consistency
 270 diagnostics are reported in Appendix Table A3.

271 **2.7 Auxiliary datasets**

272 Two auxiliary datasets are included to support spatial interpretation and benchmark curation. First,
 273 full L3 packages contain archived NOAA/NSSL FLASH CREST MaxUnitStreamflow fields.
 274 CREST is a distributed hydrological modeling framework representing runoff generation and
 275 routing processes (Wang et al., 2011), and the FLASH project applies high-resolution rainfall and
 276 hydrological modeling products to flash-flood monitoring and prediction across the United States
 277 (Gourley et al., 2017). MaxUnitStreamflow is a gridded, model-derived indicator of maximum
 278 streamflow response normalized by contributing drainage area, rather than an instantaneous
 279 streamflow observation.

280 For each full testbed package, archived MaxUnitStreamflow fields were retrieved over the package
 281 event window, clipped to the watershed domain, and stored as a Zarr array. A temporal-maximum
 282 GeoTIFF was also generated as a compact spatial summary. The released data remain gridded; the



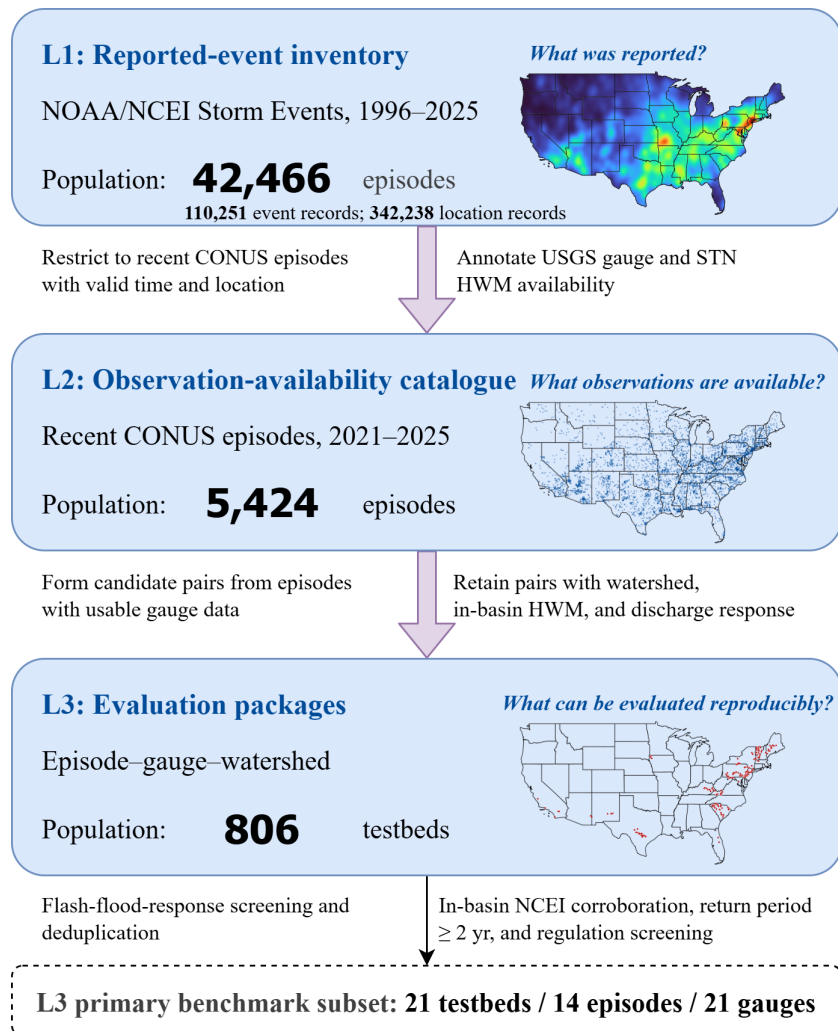
stream-network rendering used in diagnostic figures is solely a visualization procedure and does not alter the underlying raster product. MaxUnitStreamflow is included for spatial interpretation and exploratory comparison, but is not treated as an independent observation, benchmark truth, or primary-selection criterion. It is also distinct from the CREST-iMAP simulation used in the example application.

Second, wildfire perimeters were obtained from the Wildland Fire Interagency Geospatial Services (WFIGS) Interagency Fire Perimeters dataset. Incident polygons were intersected with NLDI watersheds and evaluated relative to flood-event timing to identify cases substantially affected by recent wildfire. These data support the post-selection wildfire-applicability review and associated post-fire metadata; they are not used in the initial L1–L3 qualification workflow. CREST products were accessed through the NOAA public FLASH/MRMS archives, with the Iowa State University MT Archive used where available, and WFIGS perimeters were accessed through the National Interagency Fire Center’s public geospatial services.

3 Dataset design and construction workflow

The dataset was designed to separate three related but distinct tasks: (1) preserving the national record of reported flash floods, (2) documenting whether recent reports have external observational support from gauges and high-water marks, and (3) identifying episode–gauge–watershed units that are suitable for benchmark use. We therefore organize the resource into three hierarchical levels (Figure 1; Table 2). Level 1 (L1) is a national reported-event catalogue derived from NOAA/NCEI Storm Events records for 1996–2025. Level 2 (L2) is an observation-availability candidate catalogue for recent years. Level 3 (L3) is a set of episode–gauge–watershed evaluation packages suitable for reproducible hydrological analysis, where each testbed links one flash-flood episode, one USGS stream gauge, and the gauge’s contributing watershed.

These levels answer different questions: (L1) What flash-flood episodes were reported? (L2) What external observations may be available for each recent reported episode? (L3) Which episode–gauge–watershed units are suitable for hydrological response analysis and model intercomparison? L2 is therefore not a lower-quality benchmark set; it is an observation-availability layer that retains all eligible recent episodes, including episodes for which no useful external observations are found. L3 is the level at which hydrological consistency, watershed attribution, observed discharge response, quality control, and final benchmark suitability are imposed. In this hierarchy, L1 and L2 are not intended as benchmark targets themselves; they provide provenance, observation-availability context, and a reproducible pathway for extending the dataset. The benchmark unit for model evaluation and intercomparison is L3, with the 21-testbed primary subset serving as the default conservative intercomparison set. Figure 1 summarizes the workflow and the population at each stage; Table 3 lists the contents of an L3 package.



318

319 **Figure 1. Functional organization and construction workflow of the TORRENT flash-flood**
320 **data resource.** Level 1 (L1) is a national inventory of 42,466 reported NOAA/NCEI flash-flood
321 episodes for 1996–2025, comprising 110,251 event records and 342,238 location records. Level 2
322 (L2) restricts the inventory to 5,424 recent CONUS episodes with valid temporal and spatial
323 information and annotates the availability of USGS stream-gauge observations and USGS/FEMA
324 Short-Term Network high-water marks (HWMs). Level 3 (L3) forms candidate episode–gauge pairs
325 from episodes with usable gauge data and retains 806 episode–gauge–watershed evaluation
326 packages with an available watershed, at least one in-basin HWM passing the temporal-screening
327 rule, and a detected discharge response. Subsequent flash-flood-response screening, deduplication,
328 station-relative severity screening, and post-selection curation (direct flood-control influence and
329 recent-wildfire applicability reviews) yield the primary benchmark of 21 testbeds representing 14
330 episodes and 21 gauges.



Table 2. Functional structure of the three TORRENT dataset levels. L1 preserves the national inventory and provenance of reported flash-flood episodes, L2 documents the availability of external observations for recent CONUS episodes, and L3 provides episode–gauge–watershed evaluation packages suitable for reproducible rainfall–runoff analysis. The complete L3 release contains 806 testbeds, while subsequent flash-flood-response screening, deduplication, event corroboration, severity screening, and post-selection curation define the primary benchmark of 21 testbeds. Detailed construction criteria and intermediate counts are provided in Appendix Table A1.

Level	Definition and construction	Population
L1: Reported-event inventory	NOAA/NCEI flash-flood episodes for 1996–2025, retained without hydrological screening. Includes linked event and location records for provenance and event discovery.	42,466 episodes; 110,251 event records; 342,238 location records
L2: Observation-availability catalogue	Recent CONUS episodes with valid time and location information. USGS gauge and STN HWM availability is annotated for all eligible episodes, including those without usable observations.	5,424 episodes
L3: Evaluation packages	Episode–gauge–watershed testbeds with an available watershed, at least one in-basin HWM passing the temporal-screening rule, and a detected discharge response. All packages contain response, spatial evidence, and QA/QC metadata; full packages also contain gridded forcing and auxiliary products.	806 testbeds representing 160 episodes and 208 gauges; primary benchmark: 21 testbeds representing 14 episodes and 21 gauges

338

339 3.1 Three-level dataset design

340 The workflow begins with reported flash-flood episodes rather than with stream gauges or model
341 grids, thereby preserving provenance to the national storm-event archive. A reported episode alone,
342 however, is insufficient for model evaluation: its reported location may be spatially broad, it may
343 not coincide with a usable gauge, and nearby high-water marks may represent a different flood.
344 Conversely, a stream-gauge peak alone does not establish that the response was caused by the
345 reported episode or that the reported flooding occurred within the gauge’s contributing watershed.
346 The three levels therefore serve distinct functions. L1 retains the reported flash-flood episodes
347 extracted from the NOAA/NCEI archive with minimal processing. L2 documents whether
348 potentially relevant gauge and high-water-mark observations are available while retaining episodes
349 with observation gaps. L3 creates physically interpretable benchmark testbeds only when a reported
350 episode can be linked to a specific USGS gauge, its contributing watershed, confirmed in-basin
351 flood evidence, and a detected discharge response. Each L3 testbed is defined by one episode, one
352 gauge, and one contributing watershed. A single NCEI episode may therefore produce multiple
353 testbeds when several gauges provide valid responses to the same storm; such testbeds share event
354 forcing and should not automatically be treated as statistically independent. Users seeking a
355 conservative model-intercomparison set can use the primary benchmark, whereas users requiring



larger samples can use the deduplicated or full L3 collections, or apply the released workflow to L2 episodes. Recommended uses of these collections are summarized in Sect. 5.3. The primary benchmark is a conservatively curated subset of L3 rather than a separate data level, because it retains the same episode–gauge–watershed structure while applying additional screening criteria for formal model intercomparison.

3.2 L1: National reported flash-flood episode catalog

L1 was constructed from NOAA/NCEI Storm Events records for 1996–2025. The 1996 start year was selected because it is the first year in the processed Storm Events archive for which “Flash Flood” is consistently available as an explicit EVENT_TYPE in the annual bulk records used here. The 1996–2025 period therefore provides a 30-year national archive of reported flash-flood episodes for the contiguous United States and surrounding U.S. reporting domains included in the NCEI records.

The L1 catalog is organized into three linked tables: episodes, events, and locations. An episode is the highest-level grouping used by NCEI to associate related storm reports within a common storm system. It provides the main temporal window, state and county information, Weather Forecast Office, reported casualties and damages, flood causes, and narrative description. An event is an individual Storm Events record within an episode; for flash floods, these records usually correspond to specific county-level or report-level flood occurrences with their own timing, location metadata, impacts, and descriptive information. A location record provides a geocoded reported impact point associated with an event, when such information is available. Thus, one episode may contain multiple event records, and one event may have one or more associated location points.

Annual Storm Events detail files were filtered to retain records with EVENT_TYPE = “Flash Flood”. The corresponding location files were then linked to the event-detail records to preserve the reported spatial footprint of each episode. When detailed location records were unavailable or incomplete, event-level begin and end coordinates were retained as fallback spatial information. These spatial records are not treated as hydrologic observations; they are reported locations used to support later matching with candidate stream gauges, high-water marks, and watershed boundaries.

No hydrological screening is applied at L1. In particular, L1 does not require a nearby stream gauge, high-water mark, watershed delineation, radar forcing, or observed hydrograph response. Its purpose is to preserve the full reported flash-flood population and maintain traceability to the original NCEI records. The resulting L1 catalog contains 42,466 flash-flood episodes, 110,251 event records, and 342,238 location records for 1996–2025. This catalog forms the parent population from which the L2 observation-availability catalog and the L3 benchmark testbeds are derived.

3.3 L2: Recent observation-enriched candidate catalog

L2 was designed as an inclusive observation-availability catalog for recent reported flash-flood episodes. Starting from the L1 parent catalog, episodes were retained if they occurred during 2021–2025, were located in the CONUS, had valid begin and end times, and contained at least one usable spatial coordinate from the NCEI event or location records. The 2021–2025 window was selected to balance sample size, modern data consistency, practical observation availability, and manageable quality-control effort. This recent period provides enough candidate episodes for systematic screening while reducing complications associated with older radar products, changing station



397 metadata, uneven field-observation availability, and large manual review burden.

398 For each eligible episode, L2 records the potential availability of external observations. Discharge
 399 availability was assessed with a bounded nearest-gauge probe: the NWIS instantaneous-values
 400 service was queried over the event window for up to the three nearest gauges within 25 km of the
 401 reported episode location and, when none of these returned usable discharge or stage data, for up to
 402 the three nearest gauges in the 25–50 km ring. Discharge is considered available—an episode-level
 403 flag—when any probed gauge returns at least one valid record of parameter 00060 (stage: 00065).
 404 Candidate USGS/FEMA Short-Term Network high-water marks were also identified near each
 405 episode and assigned temporal-match classes indicating whether the HWM record was confirmed,
 406 uncertain, or inconsistent with the reported episode time window. The search does not require three
 407 gauges to exist in either distance band.

408 Each episode was then assigned an observation-evidence label. A “strong” label indicates that both
 409 usable gauge data and plausible HWM support are available. A “partial” label indicates that exactly
 410 one of these observation types is available. A “weak” label indicates that nearby contextual evidence
 411 exists but is incomplete, temporally uncertain, inconsistent, or not directly usable. A “none” label
 412 indicates that no useful external observation was found. Episodes labeled “none” are retained rather
 413 than discarded, because the purpose of L2 is to document observation availability across the full
 414 recent candidate population, not to preselect benchmark cases.

415 The resulting L2 catalog contains 5,424 recent CONUS flash-flood episodes. Of these, 4,599 have
 416 discharge availability at a queried nearby gauge, 2,609 have at least one HWM within 50 km (214
 417 have at least one temporally confirmed HWM within 50 km), and the evidence-label distribution is
 418 462 strong, 4,360 partial, 431 weak, and 171 none. These labels guide downstream L3 screening,
 419 but they do not by themselves imply benchmark acceptance. Hydrological consistency, watershed
 420 attribution, in-basin HWM evidence, observed discharge response, flash-flood-response screening,
 421 and quality-control flags are applied later at L3.

422 **3.4 L3: Episode–gauge–watershed benchmark testbeds**

423 L3 converts observation-supported episodes into hydrologically interpretable testbeds, each
 424 combining a reported flood event, an observed outlet response, a watershed boundary for forcing
 425 extraction, and independent physical evidence of flooding.

426 **Candidate pairs.** The L3 workflow forms candidate episode–gauge pairs from L2 episodes that
 427 carry the discharge-availability flag and have at least one temporally confirmed HWM within 50
 428 km. For each such episode, all stream gauges within 50 km are then enumerated as possible outlets
 429 from the full upstream gauge-match table, and data usability is re-evaluated independently for every
 430 resulting episode–gauge pair in the discharge-response step below. The L2 flag, derived from the
 431 bounded probe described in Sect. 3.3, therefore acts only as an episode-level prerequisite for
 432 entering L3; in principle an episode whose only usable gauge lies outside the probed subset could
 433 be missed at this step, a conservative candidate limitation that can exclude, but never admit, episodes.
 434 The 50-km radius serves only to generate candidates and carries no hydrological meaning: relevance
 435 is established below by watershed containment, and this candidate-pair step is intentionally broader
 436 than final selection because the nearest gauge is not always the most hydrologically useful.

437 **Watershed attribution and in-basin evidence.** For each candidate gauge, the contributing



watershed is obtained from the USGS NLDI basin service, and confirmed HWMs are tested for containment within the watershed polygon. Basin retrieval was audited exhaustively; gauges genuinely absent from the NLDI index (201 sites, confirmed by repeated independent queries) are excluded at this stage. A candidate pair is retained only if at least one confirmed HWM falls inside the gauge watershed—requiring the mark to fall within the delineated watershed and pass the broad STN temporal-screening rule, rather than relying on proximity alone. Reuse of the same HWM records across multiple episode-level packages is indicated by the `hwm_qc` flag (Sect. 5.2). The NLDI watershed area is compared with the USGS-reported drainage area as a consistency check on outlet and basin delineation (median absolute difference 0.5%). This area comparison is reported as a consistency diagnostic, rather than an additional numerical L3 exclusion threshold.

Discharge response. The observed response is evaluated from the NWIS instantaneous discharge record over a window from 24 h before the episode begin time to 72 h after the episode end time. The pre-event baseline, Q_{base} , is defined as the median of all valid discharge samples during the 24 h preceding the reported episode begin time. A flood response is considered present when the hydrograph shows a clear event signal, defined by a peak-to-baseline ratio of at least 3.0, a rise from a zero (no-flow) baseline, or an absolute rise exceeding five times the pre-event standard deviation, where the standard deviation is the sample standard deviation of the same 24-h window and is computed only when more than two pre-event samples exist; with two or fewer samples it is set to zero, so this secondary criterion degenerates to requiring any positive rise. The secondary branch is intentionally permissive and affects only entry into the evaluation-ready pool; no primary testbed depends on it—all 21 primary testbeds either exceed the peak-to-baseline criterion outright (minimum ratio 4.9) or rise from a zero-flow baseline.

Assembly and tiers. A candidate pair is promoted to an evaluation-ready testbed only when three conditions hold together: an NLDI watershed is available, at least one confirmed HWM falls inside the watershed, and a discharge peak signal is detected. This produces 806 evaluation-ready testbeds across 160 episodes and 208 gauges. Each testbed is assigned a tier: Tier A (337 testbeds) has at least one NCEI flash-flood report point inside the watershed, providing spatial consistency among the reported event, the watershed, the HWM evidence, and the observed response; Tier B (469 testbeds) has in-basin HWM evidence and a discharge response but no in-basin NCEI point. Both tiers are retained; the tier is an evidence label, not an exclusion.

Flash-flood-response screen. A flash-flood screen distinguishes rapid event-scale responses from slower riverine responses using three operational criteria: rapid hydrograph rise, temporally plausible precipitation-to-peak timing, and small-to-medium watershed scale. These criteria follow the common characterization of flash floods as short-duration, rapidly responding hydrological events (Borga et al., 2011; Hapuarachchi et al., 2011; Marchi et al., 2010). The rising-limb duration is defined using an event-relative discharge threshold rather than an absolute discharge value, so that basins with different drainage areas and antecedent flows can be screened consistently. The threshold is defined as

$$Q_{thr} = Q_{base} + 0.25(Q_{peak} - Q_{base})$$

where Q_{base} is the pre-event baseline discharge as defined in the discharge-response step above (not recomputed) and Q_{peak} is the event peak discharge. The rising-limb duration is computed as the time from the last sample below the threshold immediately preceding the peak to the peak itself



(i.e., the duration of the final rise to the peak), which avoids inflating the estimate on multi-peak hydrographs and when antecedent flow is already elevated. A testbed is classified as a strict flash-flood response when the rising limb is at most 6 h, the lag between peak basin-mean precipitation and peak discharge is between -0.5 and 12 h (the small negative tolerance accommodates sub-sampling offsets between radar and gauge time stamps), and the drainage area is no greater than $2,500 \text{ km}^2$. The 6-h bound is consistent with the NWS six-hour operational timescale for flash flooding while directly requiring a rapid observed hydrograph response, rather than reproducing the NWS definition itself. This screen reduces the 806 evaluation-ready testbeds to 175 strict flash-flood testbeds (76 episodes, 68 gauges). The $2,500 \text{ km}^2$ area cutoff is an operational upper bound rather than a universal physical definition of flash flooding. One-at-a-time sensitivity of all screening thresholds is reported in Appendix Table A2; the selection is weakly sensitive to every threshold except the rising-limb bound, which is anchored to the operational timescale rather than tuned.

Deduplication. Because NCEI can split one physical storm into multiple administrative or state-level episodes, the same flood peak at the same gauge can appear more than once. Strict testbeds are therefore deduplicated by unique (gauge, peak-time) combination. Peak times are treated as identical only when they refer to the same timestamp in the retrieved NWIS instantaneous-value hydrograph; no temporal tolerance is applied, because duplicates arise when overlapping episode windows at the same gauge select the same recorded peak sample. When duplicates occur, the retained representative is selected deterministically using evidence strength (Tier A status, whether the episode window brackets the peak, the number of in-basin HWMs, the number of in-basin NCEI points, and watershed-area consistency). Deduplication reduces the strict set to 115 unique gauge–peak-time testbeds (60 NCEI episodes, 68 gauges), referred to as deduplicated testbeds hereafter.

Reported-event confirmation and flood-frequency context. An NCEI-confirmed subset is defined by requiring at least one NCEI flash-flood report point inside the watershed, ensuring direct spatial consistency with the original reported flood location; this yields 54 NCEI-confirmed testbeds (33 episodes, 46 gauges). Annual peak-flow records from NWIS are then used to place each observed discharge peak in long-term flood-frequency context. Following the Bulletin 17C distributional framework, the station annual instantaneous-peak series is fitted to a Log-Pearson Type III distribution by the method of moments with a bias-corrected station skew; the implementation uses station skew only (no regional-skew weighting or generalized-skew map) and applies a Bulletin 17B–style low-outlier test, so the resulting return periods are indicative severity estimates rather than a full regional flood-frequency analysis (England Jr et al., 2018). The distributional fit is attempted only for gauges with at least five valid annual peaks; gauges with records shorter than ten years are flagged as short-record, return periods exceeding twice the record length are flagged as extrapolations, and the record length is reported for every gauge so that highly uncertain estimates can be identified. Return period is treated as a hydrological-severity attribute rather than as the definition of flash flooding. Retaining NCEI-confirmed testbeds whose event peak corresponds to a return period of at least 2 yr—a modest station-relative severity floor that removes marginal discharge responses without restricting the benchmark to rare extremes—gives 31 testbeds (22 episodes, 26 gauges).

Post-selection curation. Two post-selection curation steps, which are scope decisions rather than tunable numerical thresholds, define the released primary benchmark. First, a direct flood-control



523 influence screening excludes gauges at or immediately downstream of flood-control structures
524 where USGS annual-peak regulation codes and structure evidence indicate that event peaks are
525 materially controlled by reservoir operations; four testbeds were excluded: two Codorus Creek
526 testbeds affected by Indian Rock Dam and Lake Marburg, the gate-controlled Canandaigua Lake
527 outlet, and the Savage River gauge immediately downstream of Savage River Dam (whose
528 unregulated upstream counterpart remains in the qualifying set). Second, a recent-wildfire
529 applicability review identifies basins substantially burned immediately before the evaluated flood
530 using WFIGS burn perimeters and event timing; the six Rio Ruidoso testbeds ($\approx 27\%$ of the
531 watershed burned two days before the first flood) are excluded from the default primary benchmark
532 because recent wildfire may materially alter infiltration and runoff generation, and standard rainfall–
533 runoff configurations lacking burn-perimeter and burn-severity inputs may therefore be
534 disadvantaged. These ten testbeds remain available in L3 with regulation and post-fire metadata.
535 The resulting primary benchmark comprises 21 testbeds representing 14 episodes and 21 gauges.
536 The complete funnel with per-stage counts, thresholds, and rationale is given in Appendix Table A1;
537 the automated stages are fully scripted and were re-executed end-to-end from the released inputs,
538 reproducing the reported membership exactly.

539 Each L3 testbed is distributed as a self-contained package (Table 3). The USGS hydrograph is the
540 benchmark response target; MRMS provides meteorological forcing in full packages; HWM and
541 NCEI points provide observation-based evidence and spatial validation context; CREST is an
542 auxiliary model-derived layer.

543



544 **Table 3.** Contents of a full L3 benchmark testbed package

File	Content
streamflow.csv	Observed USGS discharge hydrograph (cfs), native IV resolution — benchmark target
mrms_2min_precip_basin_mean.csv	Basin-mean MRMS precipitation-rate time series
mrms_2min_precip.zarr	Gridded MRMS PrecipRate over the event window
mrms_1h_rqi.zarr	Hourly MRMS Radar Quality Index (0–1) on the precipitation grid; radar sampling quality, not a calibrated error estimate
crest_maxunitstreamflow.zarr, crest_maxunitstreamflow_peak.tif	Auxiliary CREST MaxUnitStreamflow (gridded + temporal maximum)
dem.tif	3DEP terrain over the watershed domain
watershed.geojson	NLDI contributing-watershed boundary
hwms_inside.csv	In-basin STN high-water marks with quality and metadata
testbed_summary.json	Response metrics, flashiness diagnostics, return period, tier, and QA/QC flags

545

546 **3.5 Example structure of an L3 primary benchmark package**

547 Figure 2 illustrates the structure of an L3 primary benchmark testbed using the July 2025 Texas
548 flash-flood event at USGS gauge 08165500 as an example. This case is used to demonstrate the
549 information contained in each benchmark package and how the different data components support
550 interpretation of the event–gauge–watershed unit. The hierarchical directory structure and
551 constituent files of each L3 package are illustrated in Fig. A1, with detailed descriptions provided
552 in Table 3.

553 Figure 2a shows the spatial evidence layer. The NLDI watershed boundary defines the hydrological
554 unit associated with the outlet gauge, while the terrain background and stream network provide
555 physiographic context. NCEI flash-flood report locations and temporally matched HWMs are
556 plotted together with the gauge location. In this example, the watershed-scale event package
557 combines NCEI report locations with temporally matched HWMs, providing both reported-event
558 evidence and independent field evidence of flooding.



559 Figure 2b shows the primary benchmark response target: the observed USGS discharge hydrograph.
560 The hydrograph is paired with the basin-mean MRMS precipitation-rate time series to show the
561 timing of rainfall forcing and streamflow response. The event exhibits a rapid rise to a peak
562 discharge of 8,920 m³/s, demonstrating why the case satisfies the rapid-response character expected
563 of a flash-flood benchmark testbed. This panel is the main quantitative evaluation target for
564 discharge-based model intercomparison.

565 Figure 2c shows the watershed-clipped MRMS event-total precipitation field. The map summarizes
566 the spatial structure of the storm forcing over the contributing basin, with a basin-mean event total
567 of 175.5 mm and a local maximum of 313 mm. This gridded precipitation product provides the
568 meteorological forcing used to drive hydrological simulations and allows users to evaluate whether
569 the modeled response is consistent with the spatial distribution of rainfall.

570 Figure 2d shows the auxiliary CREST MaxUnitStreamflow layer sampled onto the stream network
571 for display (the released product remains gridded; Sect. 2.7). This product provides a distributed
572 model-derived representation of the event response and is included for spatial interpretation and
573 exploratory comparison. It is not treated as an independent observation or as the primary benchmark
574 target. Instead, it helps users inspect where the distributed response is concentrated relative to the
575 watershed, stream network, gauge, and observed flood evidence.

576 Together, the four panels demonstrate that an L3 benchmark testbed is not simply an event identifier
577 or a single hydrograph. It is a self-contained forcing–response–watershed–evidence package:
578 MRMS provides precipitation forcing, the USGS gauge provides the benchmark discharge response,
579 NLDI defines the hydrological unit, NCEI and HWM points provide reported and physical flood
580 evidence, 3DEP provides terrain context, and CREST provides auxiliary distributed-response
581 information. This package structure supports reproducible event-scale model evaluation while
582 retaining the spatial and observational context needed to interpret each flash-flood case.

583

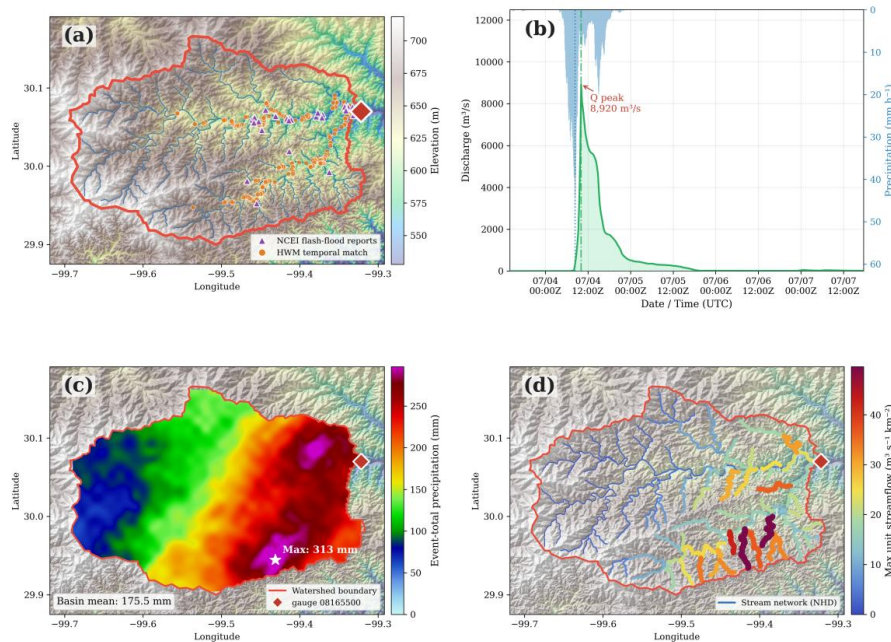


Figure 2. Example L3 primary benchmark testbed for the July 2025 Texas flash flood at USGS gauge 08165500. (a) Watershed-scale spatial evidence, including the NLDI contributing watershed, outlet gauge, NCEI flash-flood report locations, temporally matched HWMs, stream network, and terrain context. (b) Observed USGS discharge hydrograph and basin-mean MRMS precipitation-rate time series; the event peak reaches 8920 m³/s. (c) Watershed-clipped MRMS event-total precipitation, with a basin mean of 175.5 mm and a local maximum of 313 mm. (d) Auxiliary CREST MaxUnitStreamflow field, sampled onto the stream network for display only. The USGS hydrograph is the primary benchmark response target; MRMS provides forcing; NCEI and HWM points provide event evidence; and CREST is included only as an auxiliary model-derived response layer.

4 Coverage and hydrological characteristics of the primary benchmark

The primary benchmark contains 21 L3 testbeds derived from 14 flash-flood episodes and 21 USGS stream gauges (Table 4). Unless otherwise noted, all statistics in this section describe these 21 primary L3 testbeds—not the L1 inventory, the L2 candidate catalog, or the full 806-testbed L3 pool, whose populations are characterized in Sect. 3. Because the benchmark is observation-constrained, its coverage should not be interpreted as a complete climatology of U.S. flash floods. Instead, it reflects the subset of reported flash-flood episodes for which stream-gauge response, watershed attribution, field evidence, radar precipitation forcing, and quality-control criteria jointly support benchmark use. The small final sample reflects the conjunction of stringent, observation-constrained screening requirements rather than the number of flash floods occurring in 2021–2025.

4.1 Spatial, temporal, and watershed-scale coverage



607 Figure 3 summarizes the spatial, seasonal, and basin-scale distribution of the primary benchmark.
608 The primary testbeds are distributed across several regions of the contiguous United States,
609 including the Texas Hill Country and southern Plains, the Mid-Atlantic and central Appalachians,
610 the northeastern United States, and coastal California. This distribution reflects both the occurrence
611 of recent reported flash floods and the availability of co-located stream-gauge, high-water-mark,
612 and watershed evidence. The map therefore shows where high-confidence benchmark units could
613 be constructed, rather than where all flash floods occurred.

614 The 14 primary-benchmark episodes span the 2021–2025 analysis period and are concentrated in
615 the warm season, especially from late spring through early autumn. This seasonal concentration is
616 consistent with the convective rainfall environments that commonly produce short-duration flash-
617 flood response, but it is also affected by the observation-constrained screening design. The year–
618 month distribution indicates that the final benchmark includes episodes from multiple years rather
619 than a single storm season, while retaining a compact set of cases suitable for manual inspection
620 and reproducible model evaluation.

621 The primary testbeds mostly represent small to medium-sized watersheds, with drainage areas
622 ranging from approximately 100 to 2,400 km² and concentrated between roughly 200 and 800 km².
623 This scale range is appropriate for event-scale flash-flood response studies because it captures basins
624 large enough to contain stream-gauge and watershed-scale forcing information, but small enough
625 that rapid rainfall–runoff response can be meaningfully evaluated. The drainage-area distribution
626 also highlights that the benchmark is not dominated by large, slowly responding river basins,
627 consistent with the flash-flood-response screening described in Sect. 3.4.

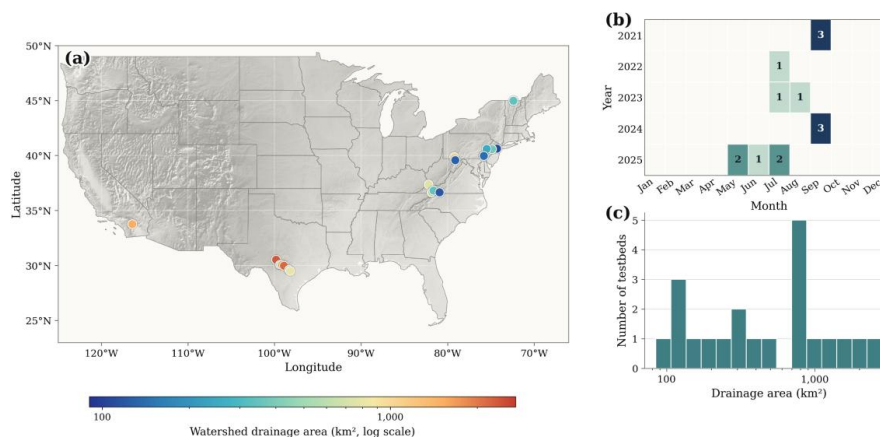


Table 4. Primary TORRENT benchmark testbeds. The primary benchmark comprises 21 episode–gauge–watershed testbeds representing 14 NCEI episodes and 21 USGS gauges, ordered by flood date. Each testbed satisfies the complete construction and screening funnel described in Appendix A1, including a retrievable watershed, in-basin high-water marks passing the temporal-screening rule, a detected gauge response, flash-flood-response screening, in-basin NCEI corroboration, and a log-Pearson type III return period of at least 2 yr. Area denotes the NLDI contributing drainage area, Q_{peak} is the observed event-peak discharge, $q_u = Q_{\text{peak}}/A$ is the unit peak discharge, RP is the estimated event-peak return period, and HWMs is the number of temporally confirmed high-water marks within the watershed.

No.	Flood date	ST	USGS gauge	Station name	Storm / event group	Area (km ²)	Peak Q (m ³ s ⁻¹)	q_u (m ³ s ⁻¹ km ⁻²)	RP (yr)	HWMs
1	2021-09-01	PA	01480500	West Branch Brandywine Creek at Coatesville ^a	Hurricane Ida remnants	119	202	1.7	24	3
2	2021-09-02	NJ	01395000	Rahway River at Rahway	Hurricane Ida remnants	106	188	1.77	79	15
3	2021-09-02	NJ	01397000	South Branch Raritan River at Stanton	Hurricane Ida remnants	381	507	1.33	97	10
4	2021-09-02	PA	01451650	Little Lehigh Creek at Tenth St. Br. at Allentown	Hurricane Ida remnants	254	215	0.85	11	9
5	2021-09-02	PA	01480617	West Branch Brandywine Creek at Modena ^a	Hurricane Ida remnants	142	348	2.45	82	6
6	2022-07-13	VA	03207800	Levisa Fork at Big Rock	SW Virginia storm, Jul 2022	769	504	0.66	5	4
7	2023-07-16	VT	04293000	Missisquoi River nr North Troy	Northeast floods, Jul 2023	339	158	0.46	3	12
8	2023-08-21	CA	10259100	Whitewater River at Rancho Mirage	Tropical Storm Hilary	1523	216	0.14	11	8
9	2024-09-27	VA	03165000	Chestnut Creek at Galax	Hurricane Helene	102	176	1.73	25	5
10	2024-09-27	VA	03471500	SF Holston River at Riverside, nr Chilhowie ^b	Hurricane Helene	198	331	1.67	>100	15
11	2024-09-27	VA	03473000	SF Holston River nr Damascus ^b	Hurricane Helene	785	748	0.95	>100	65
12	2024-09-27	VA	03474000	MF Holston River at Seven Mile Ford	Hurricane Helene	342	334	0.98	42	5
13	2025-05-13	MD	01596500	Savage River nr Barton	Central Appalachia storm, May 2025	127	233	1.83	>100	4
14	2025-05-14	PA	03079000	Casselman River at Markleton	Central Appalachia storm, May 2025	989	912	0.92	50	42
15	2025-06-12	TX	08185000	Cibolo Creek at Selma ^c	San Antonio flood, Jun 2025	710	121	0.17	2	4
16	2025-06-12	TX	08185065	Cibolo Creek nr Saint Hedwig ^c	San Antonio flood, Jun 2025	793	230	0.29	4	4
17	2025-07-04	TX	08165300	N Fork Guadalupe River nr Hunt ^d	TX Hill Country flood, Jul 2025	438	1,659	3.79	13	66
18	2025-07-04	TX	08165500	Guadalupe River at Hunt ^d	TX Hill Country flood, Jul 2025	746	8,920 †	11.96	70	183
19	2025-07-04	TX	08166200	Guadalupe River at Kerrville ^d	TX Hill Country flood, Jul 2025	1321	8,438 †	6.39	43	309
20	2025-07-04	TX	08167000	Guadalupe River at Comfort ^d	TX Hill Country flood, Jul 2025	2173	5,012	2.31	24	522
21	2025-07-13	TX	08148500	N Llano River nr Junction	TX Hill Country flood, Jul 2025	2367	1,705	0.72	10	19

Notes.

- **Storm / event groups.** Groups are meteorological classifications and may contain multiple NCEI episodes; complete episode and testbed identifiers are provided with each testbed package in the dataset.
- **Independence.** Superscripts a–d mark nested gauges on the same stream (a Brandywine; b South Fork Holston; c Cibolo Creek; d Guadalupe main stem). Aggregate evaluations should account for nested gauges and shared flood episodes, for example through episode-level aggregation or clustered resampling.
- **Data quality.** † peak determination is an approved but estimated (indirect) record at an overtopped gauge (NWIS qualifier 'e'). RP estimates above 100 yr are reported as >100 yr because extrapolation far beyond the systematic record is highly uncertain.
- **Recommended use.** The primary set is a high-confidence reference for formal model intercomparison; the strict and deduplicated tiers support robustness testing, and the full evaluation-ready pool (806 testbeds) supports exploratory analysis (Appendix A1–A2).
- **Exclusions.** Four gauges materially controlled by flood-control structures and the six-event post-fire Rio Ruidoso sequence were excluded by post-selection curation (Appendix A1, stages 12–13); they remain available in L3 with regulation and post-fire metadata.
- **Conventions.** Dates are reported in UTC. The standard evaluation window extends from 24 h before the episode beginning to 72 h after its end. Discharge files retain native NWIS units, whereas the values reported in this table are converted to SI units.



650

651 **Figure 3. Spatial, temporal, and watershed-scale coverage of the TORRENT primary**
 652 **benchmark.** (a) Geographic distribution of the 21 primary episode–gauge–watershed testbeds,
 653 plotted at their USGS gauge locations and colored by NLDI contributing drainage area on a
 654 logarithmic scale. (b) Year–month distribution of the 14 NCEI flash-flood episodes represented in
 655 the primary benchmark; numbers indicate the number of episodes in each year–month bin. (c)
 656 Distribution of contributing drainage area across the 21 primary testbeds, shown on a logarithmic x
 657 axis. Together, the panels summarize the geographic coverage, seasonal timing, and basin-scale
 658 range of the conservative primary benchmark.

659

660 4.2 Hydrological response characteristics

661 The primary benchmark testbeds represent hydrologically substantial flood responses at their
 662 corresponding stream gauges (Figure 4). The estimated return periods of the observed event peaks
 663 (Table 4) span more than two orders of magnitude, with a median value of 25 yr. This indicates that
 664 the primary benchmark subset is not composed only of marginal above-threshold events; rather, it
 665 includes many peaks that are uncommon relative to the historical annual peak-flow records at the
 666 corresponding gauges. Consistently, the observed peak-to-baseline discharge ratios span several
 667 orders of magnitude, with a median of 183 among the 19 testbeds with a nonzero pre-event baseline
 668 (Fig. 4a); the remaining two testbeds rise from a zero-flow (no-flow) baseline. Return period is used
 669 here as a station-specific severity descriptor, not as the definition of flash flooding. The flash-flood
 670 character of each testbed is instead determined by response timing, rainfall–runoff coupling,
 671 watershed scale, and observation consistency.

672 To quantify rapid hydrograph response, we use the flashiness metric (Li et al., 2023b), which
 673 measures the maximum positive increase in drainage-area-normalized discharge over a fixed
 674 duration. For a duration (d), flashiness is defined as

$$675 F_d = \frac{\max\{Q_t - Q_{t-1}, Q_t - Q_{t-2}, \dots, Q_t - Q_{t-d}\}}{A \times d}$$

676 where Q_t is the streamflow at time t , d is the duration, A is the drainage area. In Figure 4, we



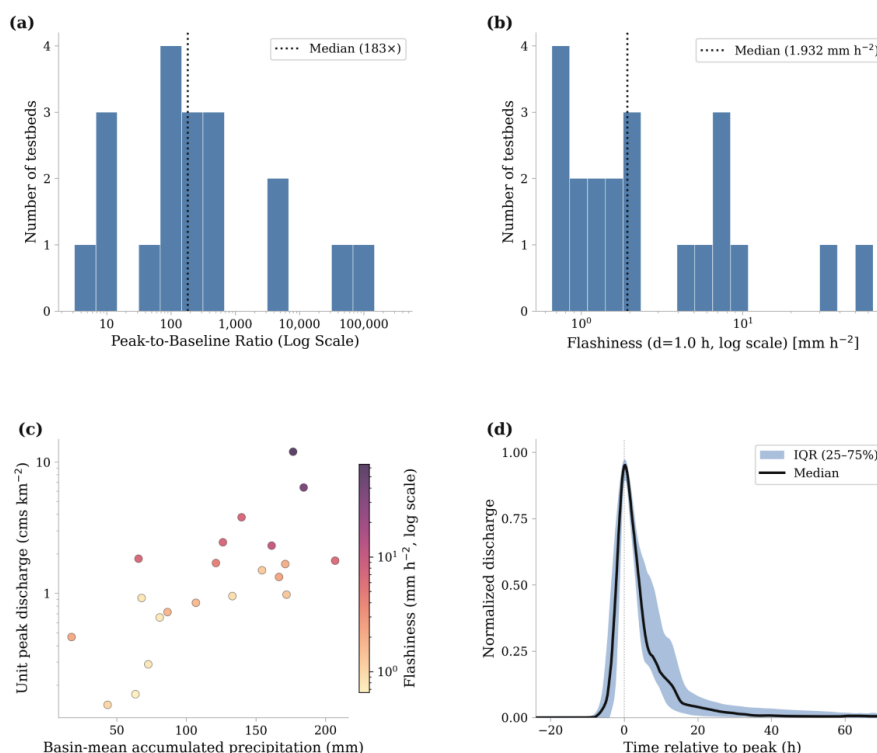
use $d = 1$ h, and F_d is reported in mm/h^2 . Larger values indicate a sharper short-duration increase in unit discharge and therefore a more rapid hydrological response. Unlike the event-level flashiness (Saharia et al., 2017), which uses the rise from action-stage discharge to peak flow, the metric used here evaluates the maximum area-normalized discharge increase over a fixed time interval.

The flashiness metric further confirms the rapid-response character of the primary testbeds. The median flashiness value is 1.932 mm/h^2 , and the distribution is concentrated on positive, high-response cases while still spanning more than one order of magnitude. The two largest flashiness values correspond to two gauge testbeds from the July 2025 Texas flash-flood event, which also appears as the example L3 benchmark package in Figure 2. These cases illustrate that the primary benchmark includes not only typical rapid-response events but also exceptionally intense, short-duration flood responses. This spread is expected for a multi-region benchmark, because flash-flood response depends on rainfall intensity, storm placement, terrain, drainage structure, and antecedent wetness. The purpose of the benchmark is therefore not to select hydrologically identical events, but to retain cases that share a rapid and observation-supported flood response while preserving meaningful hydrological variability.

The relationship between basin-mean accumulated precipitation and unit peak discharge shows that the primary testbeds include a range of rainfall forcing and runoff response magnitudes. In general, larger accumulated precipitation tends to correspond to higher unit peak discharge, but the scatter also highlights the importance of watershed-specific controls and event placement. Points colored by flashiness show that the most rapid responses are not determined by precipitation total alone; they also depend on how quickly rainfall is converted into discharge at the gauge.

The normalized hydrograph composite provides an event-centered view of response timing. After aligning each hydrograph by its observed peak and normalizing discharge by the event maximum, the median curve rises sharply near the peak and then recedes over the following hours to days. The interquartile range is narrow near the peak but broadens during the recession, indicating that the timing of peak response is consistently well defined across the primary testbeds, while post-peak recession behavior varies among watersheds. This pattern supports the use of the dataset for event-scale model evaluation focused on peak timing, peak magnitude, and early recession dynamics.

706



707

708 **Figure 4. Hydrological response characteristics of the 21 TORRENT primary benchmark**
 709 **testbeds.** (a) Distribution of the observed peak-to-baseline discharge ratio on a logarithmic scale for
 710 the 19 testbeds with a nonzero pre-event baseline; the dashed line marks the median value of 183.
 711 Two testbeds rising from a zero-flow baseline have an undefined ratio and are omitted from this
 712 panel. (b) Distribution of the 1 h flashiness metric on a logarithmic scale; the dashed line marks the
 713 median value of 1.932 mm h⁻². (c) Relationship between basin-mean accumulated precipitation and
 714 unit peak discharge, with points colored by the 1 h flashiness metric on a logarithmic scale. (d)
 715 Composite hydrograph aligned by observed peak time and normalized by the event-peak discharge;
 716 the black line shows the median and the shaded band shows the interquartile range. Together, the
 717 panels illustrate the pronounced, rapid, and hydrologically diverse discharge responses represented
 718 in the primary benchmark.

719 5 Technical validation, quality control, and recommended use

720 5.1 Quality assurance and quality control

721 Quality assurance and quality control (QA/QC) were applied at both the package and dataset levels.
 722 For each L3 testbed, automated checks verified the presence, format, temporal coverage, and
 723 internal consistency of files required for its package type. These include observed USGS discharge,
 724 watershed and HWM information, and machine-readable metadata in all packages, and MRMS
 725 forcing, RQI, auxiliary CREST products, and DEM in full packages. Key response metrics,
 726 including peak discharge, baseline discharge, peak time, peak-to-baseline ratio, flashiness



727 diagnostics, return period, and package-completeness flags, are recorded in `testbed_summary.json`.

728 A separate set of automated dataset-level validation checks was then used to re-derive the main L3
 729 qualification and selection outcomes from the released tables. These checks confirmed that each
 730 qualifying testbed has an available watershed, at least one in-basin HWM satisfying the
 731 programmed temporal-screening rule, a detected discharge response, a consistent tier assignment, a
 732 unique gauge–peak-time combination after deduplication, and a complete set of required package
 733 files. Cross-table checks also verified consistency among the testbed metadata, HWM table, NCEI
 734 location table, and per-testbed package summaries. The released benchmark passes these automated
 735 checks. In addition, as described in Sect. 3.4, the automated construction stages were re-executed
 736 end-to-end from the released inputs and reproduced the reported membership exactly.

737 Final acceptance included manual hydrological review because flash-flood evidence does not
 738 always align perfectly across reported locations, field evidence, gridded precipitation, and stream-
 739 gauge response. Manual review focused on hydrograph clarity, plausible rainfall–runoff timing,
 740 spatial consistency among NCEI reports, HWMs, gauge location, and watershed boundary, and the
 741 absence of major confounding factors such as severe data gaps, strong regulation, tidal influence,
 742 or ambiguous event attribution. The automated screening stages are re-runnable. The regulation and
 743 recent-wildfire exclusions are case-specific curation decisions summarized in Sect. 3.4 and
 744 Appendix Table A1; other manual caveats did not change primary-set membership.

745 Beyond structural validation, three physical-consistency diagnostics were computed for every
 746 primary testbed and are reported in Appendix Table A3. First, an event water-balance diagnostic
 747 compares event runoff depth with basin-mean MRMS precipitation depth. We define the apparent
 748 event runoff coefficient (RC) as $\int (Q - Q_{\text{base}})dt / (A \cdot P)$, where Q is the observed discharge,
 749 Q_{base} is the pre-event baseline discharge, A is the NLDI contributing drainage area, and P is the
 750 basin-mean accumulated MRMS precipitation depth over the precipitation accumulation window
 751 used for this diagnostic. The discharge integral uses the standard evaluation window (episode start
 752 -24 h to episode end $+72$ h). For the Missisquoi River case (USGS 04293000) alone, the
 753 precipitation depth in Table A3 was accumulated from 7 d before episode onset to 6 h after its end
 754 to capture earlier rainfall; the discharge integration window was not extended. RC lies between 0.03
 755 and 0.92 for 19 of the 21 primary testbeds. Two testbeds exceed $RC = 1$ and are flagged rather
 756 than removed because the likely causes are documented and distinct. At Hunt (08165500), the
 757 approved indirect peak estimate is retained, but the reconstructed event volume, particularly the
 758 recession limb, is inconsistent with the observed basin rainfall. This does not necessarily invalidate
 759 the approved peak estimate; the testbed remains suitable for peak-magnitude and peak-timing
 760 evaluation, whereas volume-based and water-balance metrics should be interpreted with caution. At
 761 Barton (01596500), likely underestimation of orographic rainfall by MRMS inflates the apparent
 762 runoff coefficient.

763 Second, two upper-tail diagnostics place the evaluated floods in the context of the fitted 10-yr flood:
 764 $UTR_{\text{rec}} = Q_{\text{record}}/Q_{10}$ characterizes the station flood of record, whereas $UTR_{\text{evt}} = Q_{\text{event}}/Q_{10}$
 765 characterizes the evaluated event peak (Smith et al., 2018). The median UTR_{rec} across the 21
 766 gauges is 2.26, 14 of 21 gauges have $UTR_{\text{rec}} > 2$, and the evaluated event is itself the flood of
 767 record at six gauges. These results indicate that the benchmark includes pronounced upper-tail flood
 768 responses rather than only routine high flows.



769 Third, the vertical datum of every in-basin HWM was checked. Ninety-five percent of the marks
 770 are referenced to NAVD88; original datum information is retained for the remaining marks and
 771 explicitly flagged, allowing users to restrict elevation-based comparisons to datum-consistent
 772 observations.

773 **5.2 Quality-control flags, gridded uncertainty, and evidence provenance**

774 The released metadata include quality-control flags that allow users to select testbeds appropriate
 775 for different applications. The HWM flag `hwm_qc` distinguishes packages with a distinct set of in-
 776 basin HWM records (“event_resolved”) from packages sharing the same HWM set across multiple
 777 reported episodes (“shared_primary” or “shared_borrowed”). These are organizational labels; they
 778 do not establish which individual flood produced the shared marks. This distinction is important
 779 because some USGS Short-Term Network surveys, particularly post-fire or multi-week flood
 780 campaigns, may document a broader flood period rather than a single hydrograph peak.

781 The `hwm_qc` flag is non-destructive: all original HWM records and counts are preserved, while
 782 users may apply stricter filtering for stage, elevation, or inundation diagnostics. The flag primarily
 783 affects HWM-based applications and does not alter the principal benchmark response target,
 784 because the USGS discharge hydrograph and selected peak time remain specific to each testbed.
 785 Original HWM quality grades, event associations, surveyed elevations, and vertical-datum
 786 information are retained in the released tables.

787 Additional metadata fields document other potential sources of uncertainty. The difference between
 788 NLDI-derived and USGS-reported drainage area provides a check for possible watershed or outlet
 789 mismatches. Hydrograph-completeness fields indicate whether the discharge record covers the full
 790 analysis window. Return-period, record-length, and extrapolation flags identify flood-frequency
 791 estimates that require cautious interpretation. Nested-gauge, repeated-gauge, and shared-storm
 792 indicators identify cases in which testbeds have overlapping drainage areas, use the same gauge, or
 793 represent multiple gauge responses to the same storm. Such testbeds should not be treated as
 794 statistically independent when aggregate benchmark scores are calculated.

795 Regulation and recent-burn attributes preserve the post-selection curation decisions described in
 796 Sect. 3.4. The four flood-control-affected testbeds and six post-fire Rio Ruidoso testbeds remain
 797 available in L3 with these attributes set, although they are excluded from the default primary
 798 benchmark. A water-balance flag records whether the apparent event runoff coefficient indicates
 799 potential inconsistency between the observed discharge volume and MRMS precipitation forcing.
 800 This flag allows volume-sensitive analyses to exclude or separately interpret the two testbeds
 801 identified in Appendix Table A3 without removing them from the released data resource.

802 **Uncertainty of gridded products.** TORRENT does not include calibrated pixel-wise uncertainty
 803 estimates (such as error standard deviations or confidence intervals) for the MRMS precipitation,
 804 auxiliary CREST MaxUnitStreamflow, or 3DEP terrain grids. As a complementary quality indicator,
 805 each of the 691 full L3 packages contains `mrms_1h_rqi.zarr`: hourly MRMS Radar Quality Index
 806 (RQI) on the same spatial grid as its 2 min precipitation field (Zhang et al., 2016). RQI ranges from
 807 0 to 1 and reflects radar sampling conditions, including beam blockage and beam height relative to
 808 the melting layer; it is not a calibrated precipitation-error estimate. Unavailable source records for
 809 134 hourly slots are retained as NaN and tracked in the `missing_hours` metadata. Per-package basin-



scale statistics are available in qc/rqi_summary.csv. Across the 691 full packages, the median basin-mean RQI is 0.90; 55 packages from 24 episodes (primarily in northern New England and Pennsylvania) have basin-mean RQI below 0.7. Within the 21 primary testbeds, basin-mean RQI ranges from 0.71 to 1.00, with within-event variation below 0.03. RQI and event-level precipitation–runoff consistency diagnostics (Appendix Table A3) support quality-aware use of the gridded forcing but do not supply quantitative pixel-wise uncertainty bounds.

5.3 Limitations and appropriate use

TORRENT is designed for observation-constrained studies of flash-flood hydrological response. It should not be interpreted as a complete or statistically representative climatology of flash floods in the United States. Inclusion reflects not only where flash floods were reported, but also where stream-gauge observations, HWM evidence, watershed delineations, radar precipitation forcing, and sufficiently detailed event metadata were jointly available. The resulting sample is therefore intentionally selective and observation constrained.

Several limitations should be considered. First, the 21 primary testbeds are not fully independent statistical samples. Some primary testbeds have nested watersheds or share the same physical storm, while repeated gauges also occur in the broader L3 collection. Aggregate model scores should therefore account for dependence through gauge-, watershed-, episode-, or storm-group aggregation, clustered resampling, or equivalent dependence-aware methods.

Second, HWMs provide independent physical evidence of flooding but are not continuous hydrographs and may not always be resolved to the exact time of the selected discharge peak. Their suitability for stage or inundation evaluation depends on the associated quality, survey-event attribution, and vertical-datum metadata. In particular, the June 2025 Cibolo Creek testbeds (USGS 08185000 and 08185065) are associated with HWM records from a July 2025 STN flood event, and attribution is also uncertain for the 13 July 2025 North Llano River case (USGS 08148500). These cases remain in the discharge-response benchmark, but their HWM records should not be taken as independent confirmation of the specific gauge peaks without further event-level verification. NCEI reports may also contain spatial or temporal imprecision and should be interpreted as reported-event evidence rather than exact representations of flood extent.

Third, uncertainty remains in both the observed response and meteorological forcing. Recent NWIS discharge records may be provisional, and extreme-flow values may be approved but estimated or indirectly determined when gauges are overtopped, damaged, or otherwise unable to measure the peak directly. MRMS precipitation estimates may be uncertain during extreme rainfall, at long radar range, or in complex terrain. These limitations affect derived water-balance and response metrics and are characterized through the RQI layer (Sect. 5.2), the released quality-control fields, and the diagnostics in Appendix Table A3; these are not calibrated error bounds.

Fourth, primary-set membership depends on the operational screening thresholds described in Sect. 3.4. One-at-a-time re-execution of the full construction workflow, with post-selection curation reapplied, shows that membership is less sensitive to the peak-to-baseline, precipitation-lag, drainage-area, and return-period thresholds than to the rising-limb bound. Rising-limb limits of 4, 6, 8, and 12 h produce 12, 21, 31, and 46 primary testbeds, respectively. The default 6 h criterion is anchored to the operational flash-flood timescale rather than tuned to maximize sample size or



851 model performance. Users requiring stricter or broader samples can regenerate alternative selections
 852 using the released workflow and the sensitivity results in Appendix Table A2.

853 Derived quantities should be interpreted as diagnostic descriptors rather than absolute truths. Return
 854 period provides station-relative flood-severity context, but large estimates may require extrapolation
 855 beyond the available annual peak-flow record. Peak-to-baseline ratio characterizes response contrast
 856 but can be sensitive to near-zero antecedent discharge. Precipitation-to-peak lag depends on the
 857 definition of basin-mean precipitation, the selected event window, and the temporal sampling of
 858 radar and gauge observations. Apparent event runoff coefficients inherit uncertainty from both
 859 discharge and precipitation. Where the water-balance flag is set, peak magnitude, peak timing, and
 860 rising-limb behavior are generally more defensible evaluation targets than total event volume.

861 The 21-testbed primary benchmark is recommended for controlled rainfall–runoff model
 862 intercomparison and detailed event-scale discharge-response analysis. The 115-testbed deduplicated
 863 collection supports larger-sample evaluation, but it still contains shared storms, repeated gauges,
 864 and nested watersheds; dependence should therefore be considered when aggregating results. The
 865 full 806-testbed L3 collection is most appropriate for exploratory analysis, robustness testing, and
 866 user-defined screening. L2 supports event discovery and assessment of observation availability
 867 rather than direct model benchmarking, while L1 preserves the national reported-event inventory
 868 and provenance. This tiered design lets users trade evidentiary strength against sample size while
 869 preserving full traceability of the screening process, following the data-acquisition framework
 870 articulated for prediction in ungauged basins (Hrachowitz et al., 2013).

871 For L3 rainfall–runoff modeling with full packages, the USGS discharge hydrograph is the principal
 872 response target and MRMS provides precipitation forcing. HWMs and NCEI report points provide
 873 complementary physical and reported-event evidence and may support spatial-response or
 874 inundation diagnostics when their quality-control attributes are appropriate. CREST
 875 MaxUnitStreamflow is included only as an auxiliary model-derived layer for visualization, spatial
 876 interpretation, and exploratory comparison; it should not be treated as an independent observation
 877 or benchmark truth.

878 **6 Example applications**

879 This section illustrates how a primary L3 benchmark package can support event-scale model
 880 diagnostics. The simulation is provided solely as a demonstration and is not used to determine
 881 benchmark acceptance or dataset membership. The example uses the July 2025 Texas flash-flood
 882 testbed at USGS gauge 08165500, previously shown in Figure 2. Figure 5a evaluates outlet
 883 discharge simulated by the CREST hydrological model, whereas Figure 5b demonstrates an optional
 884 spatial diagnostic from CREST-iMAP, a coupled CREST–ANUGA hydrologic–hydraulic
 885 framework (Li et al., 2021).

886 The framework couples the distributed CREST rainfall–runoff model with the ANUGA
 887 unstructured-mesh shallow-water solver. The two components are evaluated separately here. The
 888 CREST hydrological component provides the simulated outlet hydrograph shown in Figure 5a. The
 889 coupled hydraulic component routes the distributed runoff over the terrain and provides the
 890 maximum inundation extent and water-surface elevations shown in Figure 5b. The model domain
 891 covers the contributing watershed of USGS gauge 08165500 on an unstructured triangular mesh in



UTM zone 14N constructed from 10 m terrain data. The displayed configuration uses a stream-burned DEM, separate channel and floodplain friction parameters, spatially distributed CREST parameter fields, and initial soil moisture obtained from a pre-event warm-up. The simulation is forced with the same 2 min MRMS precipitation used in the benchmark package. The discharge comparison covers 3–6 July 2025 UTC, while the hydraulic simulation covers 3–5 July and brackets the observed flood peak.

Figure 5a compares the observed USGS discharge hydrograph with the CREST-simulated outlet discharge, with basin-mean MRMS precipitation shown for event context. The simulated peak is 95.0 % of the observed peak and occurs 27 min earlier than the selected observed peak (Table 5). The 1 h flashiness ratio is 0.517, indicating that CREST underestimates the sharpest short-duration rise, whereas the 6 h flashiness ratio is 0.950, indicating closer agreement over the broader multi-hour response.

The water-balance flag for this testbed should be considered when interpreting the hydrograph comparison (Sect. 5.1; Appendix Table A3). The observed peak at USGS gauge 08165500 is an approved indirect estimate, but the reconstructed event volume, particularly the recession limb, is inconsistent with the basin-mean MRMS precipitation and produces an apparent runoff coefficient greater than 1. This inconsistency does not necessarily invalidate the approved peak estimate. The comparison therefore emphasizes peak magnitude, peak timing, and rapid-rise behavior, whereas total event volume and recession-based metrics should be interpreted with caution.

Figure 5b provides a complementary spatial diagnostic using the hydraulic component of CREST-iMAP. The maximum simulated inundation extent is shown together with the watershed boundary, gauge location, and quality-screened in-basin HWMs. Because no independent observed flood-extent polygon is available for this event, the simulated inundation extent is presented for qualitative spatial interpretation rather than as a complete flood-extent validation.

The quantitative HWM diagnostic directly compares simulated and surveyed water-surface elevations. For each HWM, the maximum simulated water-surface elevation, z_{sim} , is sampled as the maximum value within a 30 m neighborhood of the reported HWM coordinate to reduce sensitivity to point-to-grid displacement. This value is compared with the surveyed HWM elevation, z_{obs} , and the signed elevation error is calculated as

$$e_z = z_{sim} - z_{obs}$$

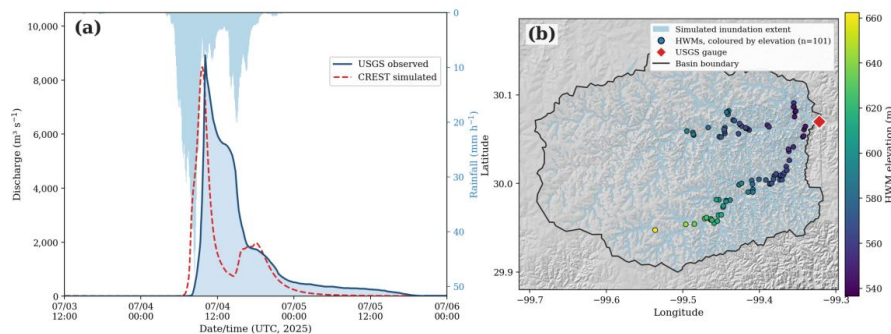
Positive errors indicate that the simulated water-surface elevation is higher than the surveyed HWM elevation, whereas negative errors indicate underestimation. The 30 m neighborhood rule is a spatial-matching choice and does not imply that the HWM observations have a uniform 30 m positional uncertainty. The comparison is therefore interpreted as a point-based water-surface-elevation diagnostic rather than a complete validation of the simulated inundation field.

Each HWM retains its original quality label. The quality-screened subset contains marks classified as Excellent, Good, or Fair that also pass the geometric-validity checks required for comparison, including a valid simulated value and a location outside the model-domain edge. This subset contains 101 marks: 8 Excellent, 61 Good, and 32 Fair. Poor and Very Poor marks are excluded from the quality-screened diagnostic but retained in the all-HWM comparison for transparency, giving 151 marks in total. For the quality-screened subset, the mean signed water-surface-elevation



error is $+0.16\text{m}$ and the median signed error is -0.07 m . For all HWMs, the corresponding mean and median errors are -0.20 and -0.12 m , respectively (Table 5). These central-error statistics indicate limited overall bias but do not establish complete spatial accuracy. Point-scale differences remain sensitive to HWM survey quality, vertical-datum consistency, terrain representation, unresolved channel geometry, local hydraulic controls, and the neighborhood-sampling procedure. Similar HWM-based hydraulic benchmarks emphasize the importance of observation quality and terrain representation when interpreting point water-level errors (Godet et al., 2025).

Together, the two diagnostics illustrate complementary uses of an L3 benchmark package. The USGS hydrograph provides the primary event-specific target for evaluating CREST-simulated discharge, including peak magnitude, timing, and rapid-rise behavior. The HWMs provide independent physical evidence for an optional point-based assessment of water-surface elevations simulated by the hydraulic component of CREST-iMAP. This example demonstrates the evaluation pathways supported by the package rather than providing a comprehensive assessment of either model component.



947

948 **Figure 5. Example application of a primary L3 benchmark testbed for the July 2025 Texas**
 949 **flash flood at USGS gauge 08165500.** (a) Comparison of observed USGS discharge and outlet
 950 discharge simulated by the CREST hydrological model, with basin-mean MRMS precipitation rate
 951 shown on the secondary axis for event context. (b) Spatial diagnostic showing the maximum
 952 inundation extent simulated by the hydraulic component of CREST-iMAP, together with the
 953 watershed boundary, USGS gauge location, and quality-screened high-water marks ($N = 101$)
 954 colored by surveyed water-surface elevation. The inundation extent is shown qualitatively, whereas
 955 Table 5 reports point-based differences between simulated maximum water-surface elevations and
 956 surveyed HWM elevations.

957

958



959 **Table 5. Example diagnostic metrics for the July 2025 Texas benchmark testbed at USGS gauge**
960 **08165500.** Streamflow metrics evaluate outlet discharge simulated by the CREST hydrological model,
961 whereas HWM metrics evaluate water-surface elevations simulated by the hydraulic component of
962 CREST-iMAP. HWM elevation errors are defined as simulated minus surveyed water-surface elevation;
963 positive values indicate overestimation. The quality-screened subset contains 101 Excellent-to-Fair
964 HWMs, and the all-HWM comparison contains 151 marks.

Component	Metric	Definition	Value	Range (Perfect)
Streamflow	Peak-discharge ratio	$\frac{Q_{peak,sim}}{Q_{peak,obs}}$	0.95	$[0, +\infty](1)$
Streamflow	Peak-timing difference	$t_{peak,sim} - t_{peak,obs}$	-0.45 hr	$[-\infty, +\infty](0)$
Streamflow	1 h flashiness ratio	$\frac{F_{1h,sim}}{F_{1h,obs}}$	0.517	$[0, +\infty](1)$
Streamflow	6 h flashiness ratio	$\frac{F_{6h,sim}}{F_{6h,obs}}$	0.950	$[0, +\infty](1)$
HWM diagnostic	Quality-screened HWM mean elevation error	$\text{mean}(z_{sim} - z_{obs})$ N = Screened HWMs	+0.16 m N=101	$[-\infty, +\infty](0)$
HWM diagnostic	Quality-screened HWM median elevation error	$\text{median}(z_{sim} - z_{obs})$ N = Screened HWMs	-0.07 m N=101	$[-\infty, +\infty](0)$
HWM diagnostic	All-HWM mean elevation error	$\text{mean}(z_{sim} - z_{obs})$ N = ALL HWMs	-0.20 m N=151	$[-\infty, +\infty](0)$
HWM diagnostic	All-HWM median elevation error	$\text{median}(z_{sim} - z_{obs})$ N = ALL HWMs	-0.12 m N=151	$[-\infty, +\infty](0)$

965

966 7 Code and Data Availability

967 The TORRENT dataset (version 1.1) and the code used to construct it are openly available from
968 Zenodo at <https://doi.org/10.5281/zenodo.23359069> (Yan et al., 2026). A standalone sample



archive (TORRENT_v1.1_sample_08165500.zip) for the July 2025 Texas flood case presented in Sect. 6 is also provided, allowing users to explore the data structure without downloading the full dataset.

A mirror of the dataset is also maintained at the Hugging Face repository `skyan1002/flash-flood-benchmark-data` (<https://huggingface.co/datasets/skyan1002/flash-flood-benchmark-data>), which provides a column-level data viewer for browsing the catalogue tables, and an interactive portal for exploring individual testbeds is available at <https://huggingface.co/spaces/skyan1002/torrent-flash-flood-agent>. The Zenodo archive is the citable version of record; the Hugging Face repository is provided for convenience and may be updated between releases.

8 Conclusions and outlook

TORRENT converts fragmented flash-flood records into traceable episode–gauge–watershed testbeds for reproducible hydrological-response analysis and model intercomparison. It links NOAA/NCEI storm reports, USGS discharge observations, MRMS precipitation, NLDI watershed boundaries, STN high-water marks, terrain data, and auxiliary model-derived products within a common package structure. Each testbed preserves the connection from the reported episode to its contributing watershed, observed gauge response and independent flood evidence, with precipitation forcing supplied in full packages, together with machine-readable metadata and quality-control information.

The three-level design separates distinct data functions rather than treating them as progressively better versions of the same benchmark. L1 preserves the national inventory and provenance of 42,466 reported flash-flood episodes for 1996–2025. L2 documents external observation availability for 5,424 recent CONUS episodes during 2021–2025, including episodes for which no useful nearby gauge or HWM evidence was found. L3 provides 806 episode–gauge–watershed evaluation packages representing 160 episodes and 208 gauges. Reproducible flash-flood-response screening, deduplication, in-basin event corroboration, station-relative severity screening, and documented post-selection curation yield a conservative primary benchmark of 21 testbeds representing 14 episodes and 21 gauges. This primary set is intended as a high-confidence reference for controlled rainfall–runoff model intercomparison, not as a statistically representative sample of CONUS flash floods. The broader L3 collection supports exploratory analysis, sensitivity testing, and user-defined screening.

The USGS discharge hydrographs, radar precipitation, independent high-water marks, and associated QC metadata allow modelers to select fit-for-purpose evaluation cases. The 21 primary testbeds span station-relative peak return periods from approximately 2 to more than 100 years; RQI layers and event-level diagnostics help identify circumstances in which input quality warrants caution. Because some cases share storms or nested watersheds, aggregate comparisons should account for statistical dependence.

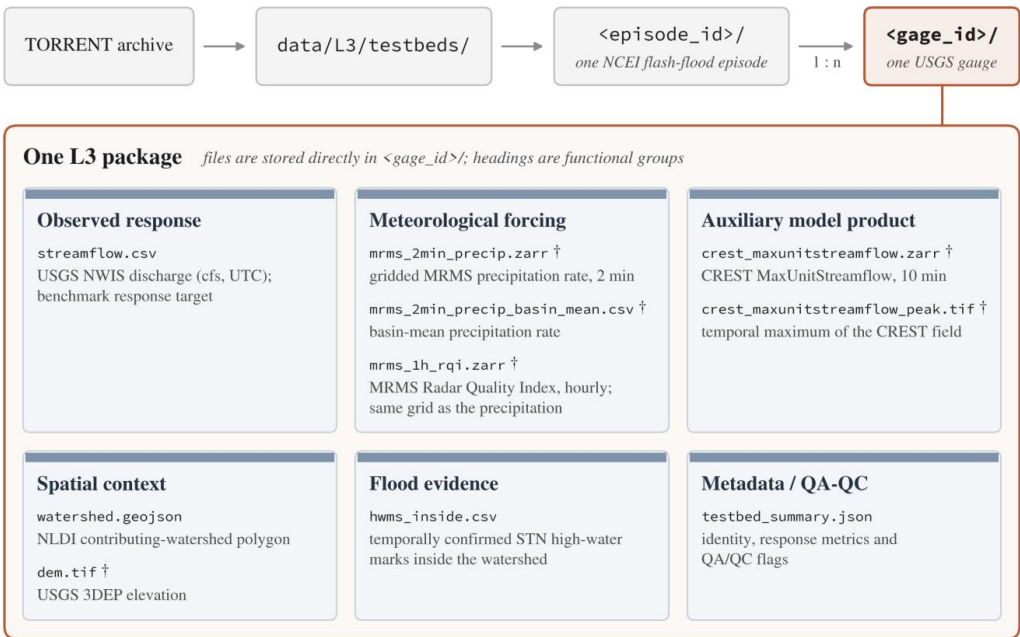
The automated construction, screening, and validation stages are released as scripted, re-runnable code, while post-selection curation decisions are explicitly documented. Future versions can incorporate additional years, newly available observations, improved event attribution, independent flood-extent information, and refined uncertainty descriptors without changing the distinction among reported-event provenance, observation availability, evaluation readiness, and primary



1010 benchmark suitability. TORRENT therefore provides both an initial CONUS benchmark for event-
1011 scale flash-flood hydrological response and an extensible framework for community development
1012 and reproducible model evaluation.



1013 **Appendix A**



† Included in full packages only.

Figure A1. Hierarchical directory structure and constituent components of an L3 testbed package in TORRENT. Each NCEI flash-flood episode may contain multiple USGS gauge-specific packages. Files are grouped by function for illustration but are stored directly within each gauge directory. The dagger (†) indicates files included only in the 691 full packages; 115 light packages retain the core observed-response, watershed, HWM, and metadata files. RQI is an hourly radar-sampling quality indicator, not a calibrated rainfall-error estimate.



Table A1. The TORRENT construction and screening funnel: criteria, thresholds, and resulting populations at each stage, from the NCEI inventory to the 21-testbed primary benchmark. Green rows annotate without exclusion; yellow rows are post-selection curation.

Each row is one stage. Filtering stages shrink the population; annotation stages add labels without exclusion (shaded green). Every removed item remains retrievable at the level where it stopped. ep = episodes; pairs = candidate (episode, gauge) combinations; tb = testbeds (episode–gauge–watershed).

#	Stage	Criterion / thresholds	Consequence (counts)	Rationale
1	L1 inventory (no screening)	All NOAA/NCEI Storm Events episodes with EVENT_TYPE = Flash Flood, 1996–2025	42,466 episodes (110,251 event records; 342,238 location points)	Authoritative national report base; preserved in full for context and provenance.
2	L2 eligibility filter	Year 2021–2025; centroid inside CONUS (lat 24–50° N, lon 125–65° W); valid begin/end time; ≥1 valid coordinate	42,466 → 5,424 ep	A recent, internally homogeneous period: stable MRMS era, active STN campaigns, retrievable NWIS records, tractable QC.
3	L2 observation annotation (no exclusion)	NWIS-IV availability (00060/00065; begin –24 h ... end +72 h) assessed using up to the three nearest candidate gauges within 50 km (primary ≤25 km, fallback 25–50 km); STN HWMs ≤50 km with temporal classes confirmed/uncertain/inconsistent → labels strong/partial/weak/none	5,424 ep labelled: strong 462, partial 4,360, weak 431, none 171 — nothing removed	Documents what CAN be observed for the full population; users trade evidentiary strength against sample size with full traceability.
4	L3 entry pre-screen	Episode has usable discharge AND ≥1 confirmed-event HWM within 50 km; episodes passing the entry gate are then expanded to all candidate gauges within 50 km of the episode report points	4,599 ep with discharge ∩ 214 with a confirmed HWM ≤50 km → 203 ep eligible → 13,315 candidate pairs	Computationally cheap pre-screen; the 50-km radius only GENERATES candidates — hydrological relevance is decided in-basin (stages 5–7).
5	Watershed delineation	USGS NLDI upstream basin polygon retrievable for the gauge (retrieval audited exhaustively; transient service failures resolved)	13,315 → 12,635 pairs (–680)	The watershed is the evaluation unit. The 680 removed pairs correspond to 201 sites confirmed absent from the NLDI index (HTTP 404 in two independent queries).
6	In-basin confirmed HWM	≥1 STN HWM with confirmed temporal match INSIDE the watershed polygon	12,635 → 1,556 pairs (–11,079)	Mapped flood evidence inside the evaluated basin meeting the stated temporal screen; individual peak attribution requires separate verification.
7	Gauged flood response	Peak/baseline ≥3.0 (baseline = median of the 24 h pre-event discharge), OR rise-from-zero, OR rise >5σ of pre-event values; window begin –24 h ... end +72 h	1,556 → 806 qualifying testbeds (160 ep, 208 gauges) (–750)	A measurable hydrograph response is required for quantitative evaluation.
8	Tier label (annotation)	Tier A: ≥1 NCEI report point inside the basin (337); Tier B: none (469)	806 tb labelled — nothing removed	Stratifies corroboration strength without discarding valid Tier-B floods.
9	Flash-flood screening	Rising limb ≤6 h (from crossing of baseline +0.25·(peak–baseline) to peak); MRMS precipitation-to-peak lag –0.5 ... 12 h; drainage area ≤2,500 km ²	806 → 175 strict tb (–631; retained with riverine/marginal class labels)	Consistent with the NWS six-hour operational timescale while directly requiring a rapid observed hydrograph response; temporally plausible precipitation–response coupling; flash-scale basins.
10	Deduplication	One testbed per (gauge, flood peak); keep-rule: Tier A > peak inside episode window > more in-basin HWMs > more NCEI points > closer NLDI/USGS area match	175 → 115 tb	The same physical flood, reported under overlapping episodes, is counted once.
11	Corroboration + severity	Tier A required; LP3 return period of the event peak ≥2 yr (station skew, Grubbs–Beck low-outlier censoring, ≥5-yr record)	115 → 54 (Tier A) → 31 tb	The final automated selection requires independent in-basin corroboration and a modest station-relative severity floor that removes marginal discharge responses.
12	Direct flood-control influence screening	Exclude gauges at or immediately downstream of flood-control structures when USGS regulation codes and structure evidence indicate event peaks are materially controlled by reservoir operations	31 → 27 tb; four regulated testbeds were excluded (Codorus ×2, Canandaigua Lake outlet, Savage below-dam). The upstream unregulated Barton gauge was already present in the qualifying set.	Peaks controlled by flood-management operations do not test rainfall–runoff skill.
13	Recent-wildfire applicability review	Identify basins substantially affected by wildfire before the evaluated flood using WFIGS burn perimeters and event timing (Rio Ruidoso: ≈27% burned two days before the first flood; six events)	27 → 21 primary testbeds (= 21 gauges) / 14 episodes; the six post-fire testbeds remain available in L3 with post-fire metadata	Recent wildfire may materially alter infiltration and runoff generation; standard rainfall–runoff configurations lacking burn-perimeter and burn-severity inputs may therefore be disadvantaged.

The automated stages (2–11) are fully scripted and were independently re-executed end-to-end; the membership reproduces exactly. Green rows indicate inventory or annotation stages without exclusion; yellow rows indicate post-selection curation and applicability review.



Table A2. One-at-a-time sensitivity of the automated screening thresholds, re-run on the full qualifying pool (806 testbeds) with the curation steps re-applied unchanged. Overlap and Jaccard are computed against the default primary benchmark ($n = 21$) using episode–gauge identities.

Each threshold is perturbed one at a time and the full automated chain (flash screen → deduplication → Tier A → return period) is re-run on the complete qualifying pool (806 testbeds); the two post-selection curation steps (A1 stages 12–13), which are scope decisions rather than tunable numerical thresholds, are re-applied unchanged after every perturbation. Results are reported against the 21-testbed primary benchmark; set overlap and Jaccard similarity are computed using episode–gauge testbed identities. All 21 primary members are retained under the relaxed settings tested. Tightened settings remove varying numbers of members, with the largest effect produced by the 4 h rising-limb threshold.

Parameter (default)	Tested	Final primary n	Overlap with default primary ($n = 21$)	Jaccard	Justification of the default
Rising-limb duration (≤ 6 h)	4 h	12	12	0.57	Rising-limb duration is the most influential tested threshold. The 6 h default is consistent with the NWS operational timescale for flash flooding while directly enforcing a rapid gauge response. Tightening to 4 h excludes nine primary members; longer thresholds progressively admit slower-response cases.
	6 h	21	21	1.00	
	8 h	31	21	0.68	
	12 h	46	21	0.46	
Precipitation-to-peak lag ($-0.5 \dots 12$ h)	8 h	18	18	0.86	Weakly sensitive. 12 h generously covers flash-scale concentration times while excluding day-scale riverine routing; the -0.5 h lower bound absorbs MRMS↔NWIS timestamp offsets.
	12 h	21	21	1.00	
	18 h	23	21	0.91	
	24 h	23	21	0.91	
Drainage area ($\leq 2,500$ km ²)	1,500	18	18	0.86	Weakly sensitive. Selected to exclude large, slow-response river basins while retaining small-to-medium flash-flood basins; doubling it adds two members.
	2,500	21	21	1.00	
	5,000 km ²	23	21	0.91	
Peak/baseline ratio (≥ 3.0)	3.0	21	21	1.00	Operates at the qualification stage (A1 stage 7). Raising the threshold to 5 changes the episode attribution of one gauge–flood case during deduplication rather than altering its physical-event coverage: the resulting set contains 21 testbeds, of which 20 share the same episode–gauge identity as the default set. 3.0 marks an unambiguous flood signal above baseline noise.
	5.0	21	20	0.91	
LP3 return period (≥ 2 yr)	1.5	23	21	0.91	Provides a modest station-relative severity floor that excludes marginal peaks without restricting the benchmark to rare extremes.
	5 yr	18	18	0.86	

Candidate-generation parameters not treated as hydrological thresholds

The 50 km gauge and HWM search radii are computational search buffers rather than benchmark-selection thresholds: final qualification requires watershed containment and therefore supersedes proximity alone. In practice the gauge radius is non-binding — primary-set gauges lie within ≈ 36 km of their episode's report points (median ≈ 4 km), and 90% of all qualifying testbeds lie within 43 km.

Interpretation: the default configuration is not a tuned optimum but a set of externally anchored, weakly sensitive choices; the single influential parameter (rising limb) is anchored to the NWS operational timescale.



Table A3. Quality-control diagnostics for the 21 primary testbeds. P is the MRMS basin-mean event precipitation depth; R is the direct-runoff depth (constant pre-event-baseline separation); RC = R/P is the event runoff coefficient, flagged where mass consistency is questionable ($RC > 1$). UTR is the upper-tail ratio of Smith et al. — the flood of record (UTR-rec) or the benchmark event peak (UTR-evt) divided by the sample 10-yr flood from the gauge's own annual-peak record. Values are diagnostic quality-control fields, not selection criteria.

Date	Gauge	Station	P (mm)	R (mm)	RC	UTR-rec	UTR-evt	QC flag
2021-09-01	01480500	West Branch Brandywine Creek At Coatesville	121	51	0.42	1.69	1.40	
2021-09-02	01480617	West Branch Brandywine Creek At Modena	126	64	0.51	1.93	1.93	
2021-09-02	01397000	South Branch Raritan River At Stanton	167	67	0.40	1.92	1.91	
2021-09-02	01451650	Little Lehigh Creek At Tenth St. Br. At Alle	107	44	0.41	2.04	1.07	
2021-09-02	01395000	Rahway River At Rahway	207	150	0.73	2.26	2.07	record-class event
2022-07-13	03207800	Levisa Fork At Big Rock	81	27	0.33	2.39	0.76	
2023-07-16	04293000	Missisquoi River Near North Troy	100	28	0.28	1.61	0.73	
2023-08-21	10259100	Whitewater R A Rancho Mirage	44	5	0.12	4.06	1.13	
2024-09-27	03473000	S F Holston River Near Damascus	133	31	0.23	2.14	2.14	record-class event
2024-09-27	03165000	Chestnut Creek At Galax	154	57	0.37	2.46	1.39	
2024-09-27	03471500	S F Holston River At Riverside, Near Chilhow	171	74	0.43	2.81	2.81	record-class event
2024-09-27	03474000	M F Holston River At Seven Mile Ford	172	61	0.35	1.90	1.55	
2025-05-13	01596500	Savage River Near Barton	66	93	1.42	2.44	2.44	RC>1 (suspected MRMS underestimation); record-class event
2025-05-14	03079000	Casselman River At Markleton	68	62	0.92	2.46	1.59	RC>0.8
2025-06-12	08185000	Cibolo Ck At Selma	64	2	0.03	2.13	0.09	
2025-06-12	08185065	Cibolo Ck Nr Saint Hedwig	73	9	0.12	1.30	0.53	
2025-07-04	08165500	Guadalupe Rv At Hunt	177	262	1.48	3.85	5.88	RC>1 (estimated indirect peak); record-class event
2025-07-04	08165300	N Fk Guadalupe Rv Nr Hunt	140	43	0.31	3.29	1.38	
2025-07-04	08166200	Guadalupe Rv At Kerrville	184	61	0.33	3.15	4.79	record-class event
2025-07-04	08167000	Guadalupe Rv At Comfort	161	47	0.29	2.64	1.94	
2025-07-13	08148500	N Llano Rv Nr Junction	86	22	0.25	1.75	1.03	

Notes.

- **Mass consistency.** Two testbeds exceed $RC = 1$, with distinct causes. Guadalupe at Hunt (1.48): the record is an approved but estimated (indirect) determination at an overtopped gauge (NWIS qualifier 'e'), and a nested volume check is decisive — the Hunt event volume (194 Mm^3) exceeds that of the containing Kerrville basin (81 Mm^3) and the basin-precipitation ceiling ($\approx 132 Mm^3$), indicating substantial uncertainty in the reconstructed event volume, particularly the recession limb; this does not necessarily invalidate the approved indirect peak estimate. The testbed is retained for peak-magnitude and peak-timing evaluation, but volume-based and water-balance metrics should be interpreted with caution. Savage River near Barton (1.42): the gauge record is unregulated and well rated, so the excess most likely reflects MRMS orographic underestimation of basin precipitation. The Missisquoi RC uses a 7-day antecedent precipitation window ($P = 100.5 mm$); the standard event window omits multi-day antecedent rainfall for this event.
- **Low RC values.** Very low coefficients are physically plausible: karst recharge on Cibolo Creek (0.03–0.12), desert-channel transmission losses at Whitewater (0.12).
- **Upper-tail behaviour.** UTR-rec median 2.26 (14 of 21 gauges > 2) confirms the benchmark samples upper-tail-heavy, flash-flood-prone basins. Six benchmark events are the flood of record at their gauge; the July 2025 Guadalupe events reach $\approx 5\text{--}6 \times$ the 10-yr flood (UTR-evt 5.9 and 4.8 at Hunt and Kerrville), exceeding the flood of record in annual-peak files that do not yet include the 2025 water year.
- **Vertical datum.** 95% of in-basin HWMs are explicitly referenced to NAVD88, consistent with the 3DEP DEM; marks without a recorded datum are flagged in the data files.



1068 **Author contribution**

1069 Conceptualization: SY and YH. Methodology: SY, ZL, SZ and YH. Data curation: SY and AS.
 1070 Funding acquisition: YH. Writing (initial): SY. Writing (review and editing): YH, MC, JG, JZ,
 1071 GB, PB, ZL, YW, XC, CD, JT and MM. Supervision: YH, ZL.

1072 **Competing Interests**

1073 The authors declare that they have no conflict of interest.

1074 **Acknowledgments**

1075 This study is supported by the University of Oklahoma Hydrology and Water Security Program.

1076 **References**

- 1077 Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment
 1078 attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*,
 1079 21, 5293–5313, <https://doi.org/10.5194/hess-21-5293-2017>, 2017.
- 1080 Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez,
 1081 M. C., Ameli, A., and Poulin, A.: A comprehensive, multisource database for
 1082 hydrometeorological modeling of 14,425 North American watersheds, *Sci Data*, 7, 243,
 1083 <https://doi.org/10.1038/s41597-020-00583-2>, 2020.
- 1084 Bates, P.: Uneven burden of urban flooding, *Nature Sustainability*, 6, 9–10,
 1085 <https://doi.org/10.1038/s41893-022-01000-9>, 2022.
- 1086 Borga, M., Anagnostou, E. N., Blöschl, G., and Creutin, J.-D.: Flash flood forecasting, warning
 1087 and risk management: the HYDRATE project, *Environmental Science & Policy*, 14, 834–844,
 1088 <https://doi.org/10.1016/j.envsci.2011.05.017>, 2011.
- 1089 Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and
 1090 Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897
 1091 catchments in Brazil, *Earth Syst. Sci. Data*, 12, 2075–2096, [https://doi.org/10.5194/essd-12-](https://doi.org/10.5194/essd-12-2075-2020)
 1092 2075-2020, 2020.
- 1093 Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane,
 1094 R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological
 1095 time series and landscape attributes for 671 catchments in Great Britain, *Earth Syst. Sci. Data*,
 1096 12, 2459–2483, <https://doi.org/10.5194/essd-12-2459-2020>, 2020.
- 1097 Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets,
 1098 F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J.,
 1099 Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation
 1100 Experiment (MOPEX): An overview of science strategy and major results from the second and
 1101 third workshops, *Journal of Hydrology*, 320, 3–17,
 1102 <https://doi.org/10.1016/j.jhydrol.2005.07.031>, 2006.



- 1103 England Jr, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas Jr, W. O., Veilleux, A. G.,
 1104 Kiang, J. E., and Mason Jr, R. R.: Guidelines for determining flood flow frequency—Bulletin 17C,
 1105 US Geological Survey, 2018.
- 1106 Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS:
 1107 hydrometeorological time series and landscape attributes for 222 catchments in Australia,
 1108 Earth Syst. Sci. Data, 13, 3847–3867, <https://doi.org/10.5194/essd-13-3847-2021>, 2021.
- 1109 Godet, J., Nicolle, P., Hocini, N., Gaume, E., Davy, P., Pons, F., Javelle, P., Garambois, P.-A.,
 1110 Lague, D., and Payrastre, O.: Benchmark dataset for hydraulic simulations of flash floods in
 1111 the French Mediterranean region, Earth System Science Data, 17, 2963–2983,
 1112 <https://doi.org/10.5194/essd-17-2963-2025>, 2025.
- 1113 Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P.-E., Clark, R. A., Argyle, E., Arthur, A.,
 1114 Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., and Howard, K. W.: The FLASH Project:
 1115 Improving the Tools for Flash Flood Monitoring and Prediction across the United States,
 1116 <https://doi.org/10.1175/BAMS-D-15-00247.1>, 2017.
- 1117 Hapuarachchi, H. A. P., Wang, Q. J., and Pagano, T. C.: A review of advances in flash flood
 1118 forecasting, Hydrological Processes, 25, 2771–2784, <https://doi.org/10.1002/hyp.8040>, 2011.
- 1119 Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W.,
 1120 Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V.,
 1121 Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S.,
 1122 Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of
 1123 Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–
 1124 1255, <https://doi.org/10.1080/02626667.2013.803183>, 2013.
- 1125 Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology
 1126 and Environmental Sciences for Central Europe, Earth Syst. Sci. Data, 13, 4529–4565,
 1127 <https://doi.org/10.5194/essd-13-4529-2021>, 2021.
- 1128 Koenig, T. A., Bruce, J. L., O'Connor, J., McGee, B. D., Jr, R. R. H., Hollins, R., Forbes, B. T., Kohn,
 1129 M. S., Schellekens, M., Martin, Z. W., and Peppler, M. C.: Identifying and preserving high-water
 1130 mark data, Techniques and Methods, 3-A24, U.S. Geological Survey,
 1131 <https://doi.org/10.3133/tm3A24>, 2016.
- 1132 Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L.,
 1133 Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community
 1134 dataset for large-sample hydrology, Sci Data, 10, 61, <https://doi.org/10.1038/s41597-023-01975-w>, 2023.
- 1136 Li, Z., Chen, M., Gao, S., Luo, X., Gourley, J. J., Kirstetter, P., Yang, T., Kolar, R., McGovern, A.,
 1137 Wen, Y., Rao, B., Yami, T., and Hong, Y.: CREST-iMAP v1.0: A fully coupled hydrologic-
 1138 hydraulic modeling framework dedicated to flood inundation mapping and prediction,
 1139 Environmental Modelling & Software, 141, 105051,



- 1140 <https://doi.org/10.1016/j.envsoft.2021.105051>, 2021.
- 1141 Li, Z., Gao, S., Chen, M., Gourley, J. J., Liu, C., Prein, A. F., and Hong, Y.: The conterminous
1142 United States are projected to become more prone to flash floods in a high-end emissions
1143 scenario, *Commun Earth Environ*, 3, 1–9, <https://doi.org/10.1038/s43247-022-00409-6>, 2022.
- 1144 Li, Z., Xue, X., Clark, R., Vergara, H., Gourley, J., Tang, G., Shen, X., Kan, G., Zhang, K., Wang, J.,
1145 Chen, M., Gao, S., Zhang, J., Yang, T., Wen, Y., Kirstetter, P., and Hong, Y.: A decadal review of
1146 the CREST model family: Developments, applications, and outlook, *Journal of Hydrology X*,
1147 20, 100159, <https://doi.org/10.1016/j.hydroa.2023.100159>, 2023a.
- 1148 Li, Z., Gao, S., Chen, M., Zhang, J., Gourley, J. J., Wen, Y., Yang, T., and Hong, Y.: Introducing
1149 Flashiness-Intensity-Duration-Frequency (F-IDF): A New Metric to Quantify Flash Flood
1150 Intensity, *Geophysical Research Letters*, 50, e2023GL104992,
1151 <https://doi.org/10.1029/2023GL104992>, 2023b.
- 1152 Marchi, L., Borga, M., Preciso, E., and Gaume, E.: Characterisation of selected extreme flash
1153 floods in Europe and implications for flood risk management, *Journal of Hydrology*, 394, 118–
1154 133, <https://doi.org/10.1016/j.jhydrol.2010.07.017>, 2010.
- 1155 Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J. C. J. H., Bates, P., Bertola, M., Kemter,
1156 M., Kreibich, H., Lall, U., and Macdonald, E.: Causes, impacts and patterns of disastrous river
1157 floods, *Nat Rev Earth Environ*, 2, 592–609, <https://doi.org/10.1038/s43017-021-00195-3>,
1158 2021.
- 1159 Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett,
1160 D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample
1161 watershed-scale hydrometeorological data set for the contiguous USA: data set
1162 characteristics and assessment of regional variability in hydrologic model performance,
1163 *Hydrology and Earth System Sciences*, 19, 209–223, [https://doi.org/10.5194/hess-19-209-](https://doi.org/10.5194/hess-19-209-2015)
1164 2015, 2015.
- 1165 Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., Hong, Y., and Giroud, M.: Mapping Flash
1166 Flood Severity in the United States, *Journal of Hydrometeorology*, 18, 397–411,
1167 <https://doi.org/10.1175/JHM-D-16-0082.1>, 2017.
- 1168 Slater, L., Villarini, G., Archfield, S., Faulkner, D., Lamb, R., Khouakhi, A., and Yin, J.: Global
1169 Changes in 20-Year, 50-Year, and 100-Year River Floods, *Geophysical Research Letters*, 48,
1170 e2020GL091824, <https://doi.org/10.1029/2020GL091824>, 2021.
- 1171 Smith, J. A., Cox, A. A., Baeck, M. L., Yang, L., and Bates, P.: Strange Floods: The Upper Tail of
1172 Flood Peaks in the United States, *Water Resources Research*, 54, 6510–6542,
1173 <https://doi.org/10.1029/2018WR022539>, 2018.
- 1174 Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T.
1175 A., and Slayback, D. A.: Satellite imaging reveals increased proportion of population exposed



- 1176 to floods, *Nature*, 596, 80–86, <https://doi.org/10.1038/s41586-021-03695-w>, 2021.
- 1177 Wang, J., Hong, Y., Li, L., Gourley, J. J., Khan, S. I., Yilmaz, K. K., Adler, R. F., Policelli, F. S., Habib,
 1178 S., Irwin, D., Limaye, A. S., Korme, T., and Okello, L.: The coupled routing and excess storage
 1179 (CREST) distributed hydrological model, *Hydrological Sciences Journal*, 56, 84–98,
 1180 <https://doi.org/10.1080/02626667.2010.543087>, 2011.
- 1181 Wing, O. E. J., Lehman, W., Bates, P. D., Sampson, C. C., Quinn, N., Smith, A. M., Neal, J. C.,
 1182 Porter, J. R., and Kousky, C.: Inequitable patterns of US flood risk in the Anthropocene, *Nat.*
 1183 *Clim. Chang.*, 12, 156–162, <https://doi.org/10.1038/s41558-021-01265-6>, 2022.
- 1184 Yan, S., Hong, Y., and Ma, M.: TORRENT: a benchmark-ready flash-flood dataset for the
 1185 contiguous United States (1.1), <https://doi.org/10.5281/ZENODO.23359069>, 2026.
- 1186 Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S.,
 1187 Martinaitis, S., Arthur, A., Cooper, K., Brogden, J., and Kitzmiller, D.: Multi-Radar Multi-Sensor
 1188 (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities, *Bulletin of the*
 1189 *American Meteorological Society*, 97, 621–638, <https://doi.org/10.1175/BAMS-D-14-00174.1>,
 1190 2016.
- 1191 Zhu, S., Li, Z., Yan, S., Xie, H., Yu, G., Gourley, J. J., Chen, M., and Hong, Y.: Widespread
 1192 intensification of global river hydrograph flashiness under climate change, *Commun Earth*
 1193 *Environ*, <https://doi.org/10.1038/s43247-026-03681-y>, 2026.

1194