https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Outrunning flash floods: XGBoost and sparse impact reports deliver global medium-range probabilistic forecasts of flash flood occurrence

Fatima M. Pillosu [1,2], Mariana Clare [2], Calum Baugh [2], Florian Pappenberger [2], Christel Prudhomme [2], and Hannah L. Cloke [1,3]

[1]Department of Geography and Environmental Science, University of Reading, Whiteknights Campus, PO Box 227, Reading, RG6 6AB, UK
[2]European Centre for Medium-range Weather Forecasts, Shinfield Rd, Reading, RG2 9AX, UK
[3]Department of Meteorology, Brian Hoskins Building, University of Reading, Whiteknights Road, Earley Gate, Reading, RG6 6ET, UK

Correspondence: Fatima M. Pillosu (fatima.pillosu@ecmwf.int)

Abstract. Flash floods are the world's most frequent and deadly type of flood. Yet, no medium-range forecasts of their occurrence exist over a continuous global domain - essential to fulfil the UN's "Early Warnings for All" target to protect everyone with early warning systems. This study addressed this gap in two phases. In a first phase, regional medium-range, data-driven forecasts of flash occurrence were developed by combining regional high-density, quality-controlled flash flood impact reports (e.g., NOAA's Storm Event Database over the Contiguous US) with global reanalysis and forecasts (e.g. from ERA5 for non-meteorological variables and ERA5-ecPoint for rainfall). Out of all the tested models, XGBoost gradient boosting achieved the best performance: it maintained high and constant discrimination skill across scores (e.g. ROC and Precision-Recall curves) and lead times, and forecast probabilities remained reliable below 10% at day 1 and 2% at day 5. In a second phase, a spatial-constrained sensitivity analysis evaluated how well the regional XGBoost model generalised to unseen regions. The sensitivity analysis revealed that a model trained on hydro-climatologically diverse and observation-dense sub-domains generalised better than those trained across the full domain with sparser data, suggesting a viable strategy for extending regionally trained forecasts of flash flood occurrence globally. Hence, this study provides the first empirical evidence that global, medium-range forecasts of flash flood occurrence are achievable with simple data-driven approaches and readily available data, closing one of the most pressing and long-standing gaps in modern hydrology.

1 Introduction

Flash floods are emerging as one of the most pressing hazards of the 21st century (WMO, 2025). They account for approximately ~85% of flood events worldwide, represent the deadliest flood type, and cause over 5,000 fatalities annually (Dordevic et al., 2020). Recent catastrophic events - in Germany and China (2021), Libya (2023), Spain, Pakistan, Afghanistan, the UAE (2024), and the US (2025) - demonstrate that flash floods cause devastating impacts across the world. To the immediate socio-economic (Ebi et al., 2021) and environmental (Zhang et al., 2024) consequences of flash floods, long-term psychological

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table 1. Definitions of key concepts used in this study.

Concept	Definition
Fluvial flash floods	They result from the rapid rise and overbank flow of rivers, streams, or other defined watercourses following localised extreme rainfall in small ($<$500 km^2), steep catchments, with concentration times from minutes to hours (NWS, 2025).
Pluvial (urban) flash floods	They occur when rainfall intensity exceeds the capacity of local infiltration or urban drainage systems, causing surface water flooding, independent of any watercourse (Speight et al., 2021).
Medium-range forecasts	Forecasts from day 3 to day 7 (NOAA), 10 (MetOffice), or 15 (ECMWF)*.
Forecasts with a continuous global domain	Forecasts over all land grid-boxes worldwide, excluding ocean points. The domain is termed *continuous* because every land grid-box receives a forecast, in contrast to systems that cover only discrete catchments, selected reporting points, or patchy regional domains with gaps between them.
Flash flood occurrence (observation)	Binary event within a given grid-box (e.g. ERA5) and accumulation period (e.g. 24-hourly, 00-00 UTC): yes-event (1) if at least one (fluvial and pluvial) flash flood report is recorded within the grid-box during the accumulation period; non-event (0) otherwise.
Flash flood occurrence (continuous forecast)	The probability (from 0 to 100%) that a (fluvial and pluvial) flash flood event occurs within a given grid-box (e.g. ERA5) during a specified accumulation period (e.g. 24-hourly, 00-00 UTC).
Flash flood occurrence (binary forecast)	Binary event within a given grid-box (e.g. ERA5) and accumulation period (e.g. 24-hourly, 00-00 UTC): yes-event (1) if the predicted (fluvial and pluvial) flash flood probability exceeds a chosen threshold; non-event (0) otherwise.
Severe class imbalance	It refers to a training dataset in which one class is vastly outnumbered by the other one, with a minority-to-majority class ratio above the 1:100 threshold (Leevy et al., 2018)

* NOAA definition: https://forecast.weather.gov/glossary.php?letter=m; MetOffice definition: https://weather.metoffice.gov.uk/guides/10-day-forecast; ECMWF definition: https://confluence.ecmwf.int/display/FUG/Section+2.1.2.1+Medium+Range+Ensemble+forecasts

effects in affected populations may also follow (Iqbal et al., 2023). As climate change increases the frequency and intensity of extreme rainfall (IPCC, 2023), including in historically low-risk regions (Fowler et al., 2021), existing vulnerabilities are projected to escalate (Hirabayashi et al., 2021). With the UN's "Early Warnings for All" initiative (UN, 2022), the most recent international target is to provide worldwide flash flood early warning with sufficient lead time for emergency action (Bazo et al., 2019).

Despite advances in flash flood prediction at local and regional scales (Sadkou et al., 2024), significant obstacles persist in developing medium-range forecasts over a continuous global domain (see definition in Table 1), including uncertainty in localised rainfall forecasts, computational demands, and limited hydrological observations (Zanchetta and Coulibaly, 2020; Speight et al., 2021). Current physically-based, operational systems provide regional (Ibarreche et al., 2020), national (Liu et al., 2018; Javelle et al., 2016; Maybee et al., 2024), and continental coverage (Gourley et al., 2017; Flamig et al., 2020). The Global Flood Awareness System (GloFAS) provides flood predictions over a continuous global domain but cannot resolve small, flashy catchments due to data and computational constraints (Matthews et al., 2025). The European Flood Awareness System

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

(EFAS) predicts flash floods up to day 5 using ERIC, a dynamic runoff coefficient (Raynaud et al., 2015). However, these forecasts cover only Europe and parts of Northern Africa, and only for catchments with upstream drainage areas exceeding 150 km2[1], leaving smaller catchments uncovered. WMO's Flash Flood Guidance System (FFGS) attempts global coverage by implementing numerous regional forecasting systems (Georgakakos et al., 2022). Their spatial coverage remains patchy, covering only 40% of the global population[2]. Moreover, the system's reliance on high-density observational networks and km-scale NWP model outputs limits forecast lead times to a few hours up to 1 day, whilst high computational costs compromise scalability in regions with limited economic resources[3]. These persistent limitations of physically-based operational systems have prompted in recent years the exploration of alternative modelling approaches as a complementary pathway for extending the domain coverage and lead time of forecasts of flash flood occurrence (see definition in Table 1).

Machine learning (ML) approaches have demonstrated remarkable success in riverine flood prediction (Nearing et al., 2024; Kratzert et al., 2024), yet their application to flash floods remains confined to catchment or regional scales, with forecast lead times rarely exceeding 24 hours (Santos et al., 2025). Two fundamental obstacles prevent extension to global scales and up to medium-range lead times. First, the observational basis for training is severely limited: discharge observations for small catchments constitute less than 1% of large-sample datasets such as CARAVAN (Kratzert et al., 2023) and the most recent single-country CAMELS datasets (Nijzink et al., 2025; Jimenez et al., 2025; Bushra et al., 2025; Liu et al., 2025; Delaigue et al., 2025). Moreover, these records capture only fluvial flash floods, leaving pluvial flash floods entirely unrepresented (see definitions in Table 1). Flash flood impact reports offer an alternative target variable that covers both fluvial and pluvial events and directly represents the quantity of greatest operational value, i.e., flash flood occurrence, upon which forecasters and emergency managers can act more readily than on uncertain discharge predictions, particularly at medium-range lead times where hydrological uncertainty compounds. In this study, the term "flash flood" is, therefore, intended to embrace both types of flash flood. Impact reports remain sparse and geographically biased too (Panwar and Sen, 2020). Second, existing data-driven flash flood models rely on regional, km-scale NWP output, limiting both spatial coverage and forecast lead times. Global NWP models extend coverage to a continuous global domain and provide forecasts up to medium-range lead times, but their coarser resolution has historically discouraged their use for flash flood prediction (Emerton et al., 2016). Despite machine learning's capacity to handle sparse datasets (Altalhan et al., 2025), and the growing interest in using global NWP forecasts for flash flood prediction (Bucherie et al., 2022) due to their improved skill in predicting extreme rainfall (Haiden et al., 2025), also when compared against higher-resolution alternatives (Hewson, 2024), no study has yet demonstrated the feasibility of combining global, medium-range NWP forecasts with flash flood impact reports using machine learning to produce medium-range, data-driven flash flood predictions over a continuous global domain. This, therefore, remains one of the pressing challenges in modern hydrology (Al-Rawas et al., 2024).

To address this gap, this study introduces a framework to develop and evaluate data-driven, medium-range forecasts of the probability of flash flood occurrence within each grid-box, over a continuous global domain. The term *forecast* is used

[1]https://confluence.ecmwf.int/display/CEMS/ERIC+reporting+points
[2]https://wmo.int/activities/flash-flood-guidance-system-global-coverage-ffgs
[3]Information on the technical requirements of the FFGS can be found in the FFGS "Sustainability Documents" at https://community.wmo.int/events/second-ffgs-programme-management-committee-pmc-meeting-joint-pmc-eg-meeting-and-eg-meeting

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

throughout to denote predictions driven by NWP outputs at specific lead times, distinguishing the approach from static flood risk mapping or climatological susceptibility assessments. The forecast horizon was limited to day 5 because, beyond this range, the chaotic growth of atmospheric uncertainty substantially degrades the predictability of localised extreme rainfall that drives flash flood occurrence (Žagar, 2017). The study addresses two main research questions. First, can data-driven models effectively integrate hydro-meteorological variables from global reanalysis and NWP forecasts with flash-flood impact reports to generate skilful forecasts up to day 5? Second, can models trained on high-quality regional observations retain meaningful predictive capability when deployed in data-scarce regions outside their training domain to achieve global coverage? To address the first research question, we formulated the problem as a probabilistic classification of flash-flood occurrence at the grid scale of the hydro-meteorological datasets used as model predictors. The target variables consist of flash-flood impact reports from the high-resolution, regional NOAA's Storm Event Database[4] covering the CONtiguous US (CONUS). The framework is driven by hydrological, static, and climatological variables from ERA5 reanalysis and forecasts up to day 5 (Hersbach et al., 2020), and by the ERA5 rainfall reanalysis and forecasts post-processed with the ecPoint technique (Hewson and Pillosu, 2021), called ERA5-ecPoint[5]. Both datasets provide global coverage. A key feature of the framework is its reliance on parsimonious, well-understood machine-learning models, prioritising robustness, interpretability, and operational feasibility over architectural complexity. Explainable AI techniques, such as SHAP (Zhao et al., 2025a), were also employed in this study to verify that the model captures physically meaningful relationships, diagnose potential spurious correlations, and build confidence in the predictions for operational deployment. The second research question is of particular operational significance because if regionally-trained models generalise successfully, high-quality observations from well-monitored regions could underpin flash flood forecasts in areas that currently lack any warning capability. To test this, a sensitivity analysis systematically degrades and spatially restricts the training observations over the CONUS to simulate the data-scarce conditions that characterise much of the global domain.

The manuscript is organised as follows. Section 2 describes the study domain, observational database, and model features. Section 3 details the machine-learning framework, including the SHAP-based approach used to interpret model outputs, and the sensitivity analysis design for the global expansion of the regional training. Section 4 presents the verification results, whilst the section 5 illustrates model performance through case studies. Section 6 interprets the findings and their operational implications, and Section 7 summarises the main conclusions of the study.

2 Data

2.1 Study domain

The CONUS serves as the primary regional study domain. It contains diverse physiographic regions (Figure 1a). From flat, expansive terrain dominating the Great Plains in the central US, west of the 100°W meridian, and the plains along the Atlantic Coast and the Gulf of Mexico, the CONUS also contains elevations exceeding 2,000 m a.s.l. in the Appalachian Mountains

[4]see Section 2.2 for a full description of the dataset.
[5]see Section 2.3 for a full description of the datasets.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

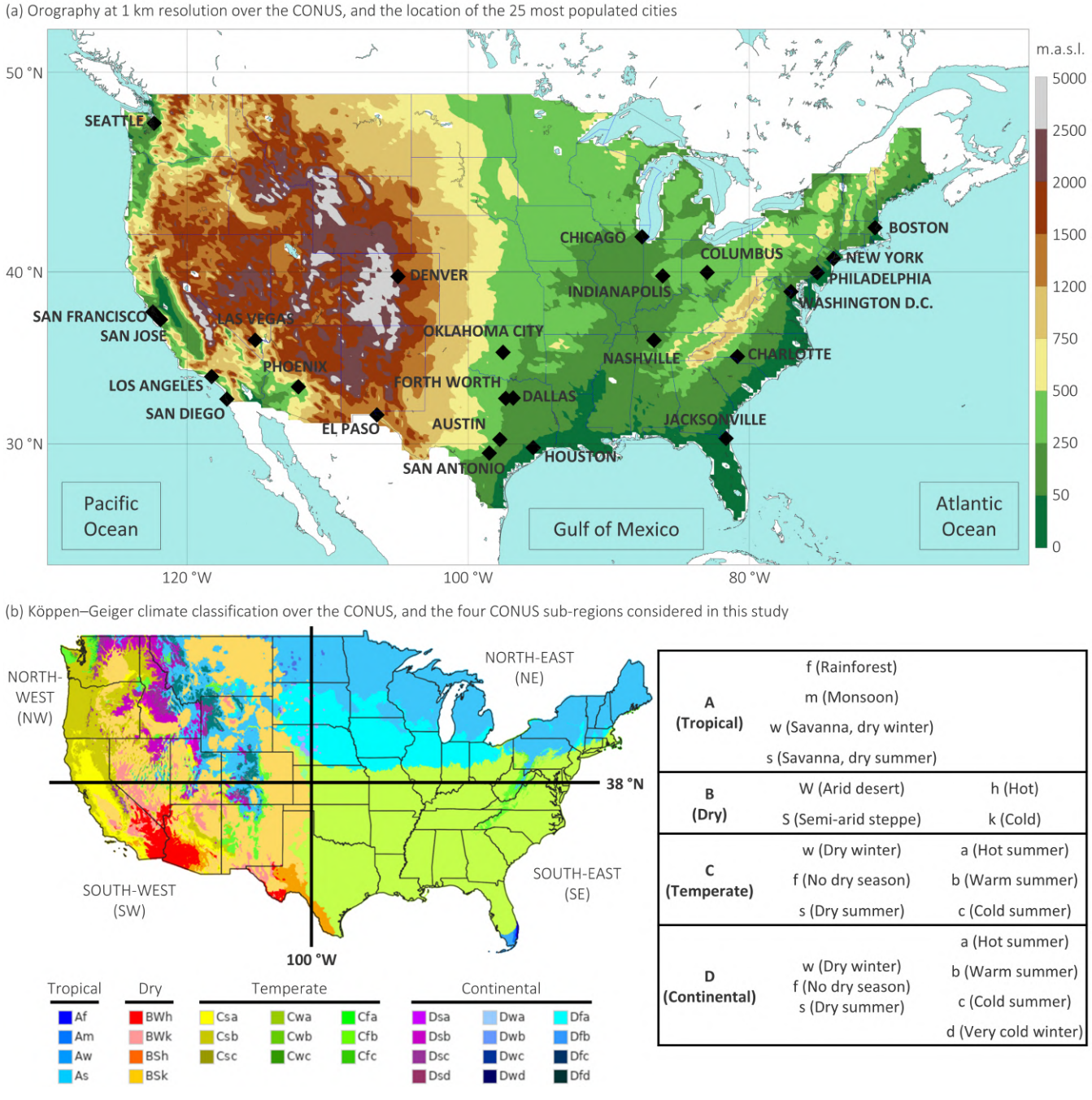

Figure 1. Study domain and climatic context over the CONUS. (a) Orography at 1 km resolution (green–brown shading) with the 25 most populated cities (black dots). (b) Köppen–Geiger climate classification (map source: en.wikipedia.org/wiki/Climate_of_the_United_States), with the four CONUS sub-regions used in this study: North-West (NW), North-East (NE), South-East (SE), and South-West (SW).

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

to the east and surpassing 4,000 m a.s.l. across the Rocky Mountains and the Sierra Nevada in the west. The Köppen–Geiger classification reveals a wide range of climatic regimes (Figure 1b). From arid and semi-arid conditions in the interior West to temperate and continental regimes across the eastern half, Mediterranean climates along the Pacific Coast, and a narrow tropical fringe in southern Florida. To facilitate regional analysis of model performance across these contrasting regimes, the domain was subdivided into four sub-regions, North-West (NW), North-East (NE), South-West (SW), and South-East (SE), delineated by the 38°N parallel and 100°W meridian (Figure 1b). Notably, the eastern half of the CONUS concentrates the majority of the 25 most populated cities (Figure 1a) and the highest population densities (not shown). This mosaic of physiographic and climatic regimes exposes the machine learning models to a broad spectrum of flash-flood-generating mechanisms, from intense convective precipitation over the Great Plains and Southeast to rain-on-snow events in mountainous catchments, land-falling hurricanes along the Gulf and the Atlantic coast, and large-scale systems such as atmospheric rivers impacting the Pacific coast (Dougherty and Rasmussen, 2019; Saharia et al., 2017). Such diversity is a prerequisite for training a model intended for global transfer.

2.2 Flash flood impact reports: NOAA's Storm Event Database

The Storm Event Database, maintained by NOAA's National Centers for Environmental Information (NCEI), is the US's official repository of severe weather (point) impact records. It covers the CONUS and overseas territories. This study considers point reports only over the CONUS because it encompasses sufficient hydro-climatic diversity for global transfer as described in Section 2.1. This study uses version 3.1, retaining point flash flood reports from 2001 to 2024 (Figure 2a); earlier records lack latitude/longitude coordinates essential for model development and verification, and 2025 records have not undergone quality control by NOAA at the time of writing.

The Storm Event Database provides extensive metadata, compiled by professional meteorologists and hydrologists, including impact severity, affected areas, and casualty information (see Table A1 in Appendix A for more details). The database provides high-density, quality-controlled impact reports, whose consistent reporting protocols mitigate the underreporting and spatial biases that characterise global datasets (Panwar and Sen, 2020; Gaume et al., 2009). However, as with any human-augmented reporting system, it is subject to known biases. Population density is one of the most significant factors influencing report location (Marjerison et al., 2016): the more densely populated eastern CONUS (as shown in Figure 1a and discussed in Section 1) yields a substantially higher frequency of reports than the west (Figure 2b). Diurnal cycles of human activity, transcription errors, and memory errors further affect report timing and location (Barthold et al., 2015). Evidence of these issues has been found in assessments of Flash Flood Guidance (FFG) skill (Clark et al., 2014). Despite these limitations, the Storm Event Database preserves fundamental hydro-climatological signals. The 2021 time series (Figure 2c, red bars for point reports counts) exhibits the expected seasonal peak during summer and early autumn (Dougherty and Rasmussen, 2019). Peak daily counts (blue circle in Figure 2c) do not necessarily reflect widespread flooding across the domain, as reports may be spatially concentrated over small areas during individual events (e.g., Storm Ida over New York, Figure 2d). Nevertheless, daily report counts are overall sufficient for robust model training and evaluation. A detailed description of the post-processing applied to the impact reports to prepare them for model training and validation is provided in Appendix A.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

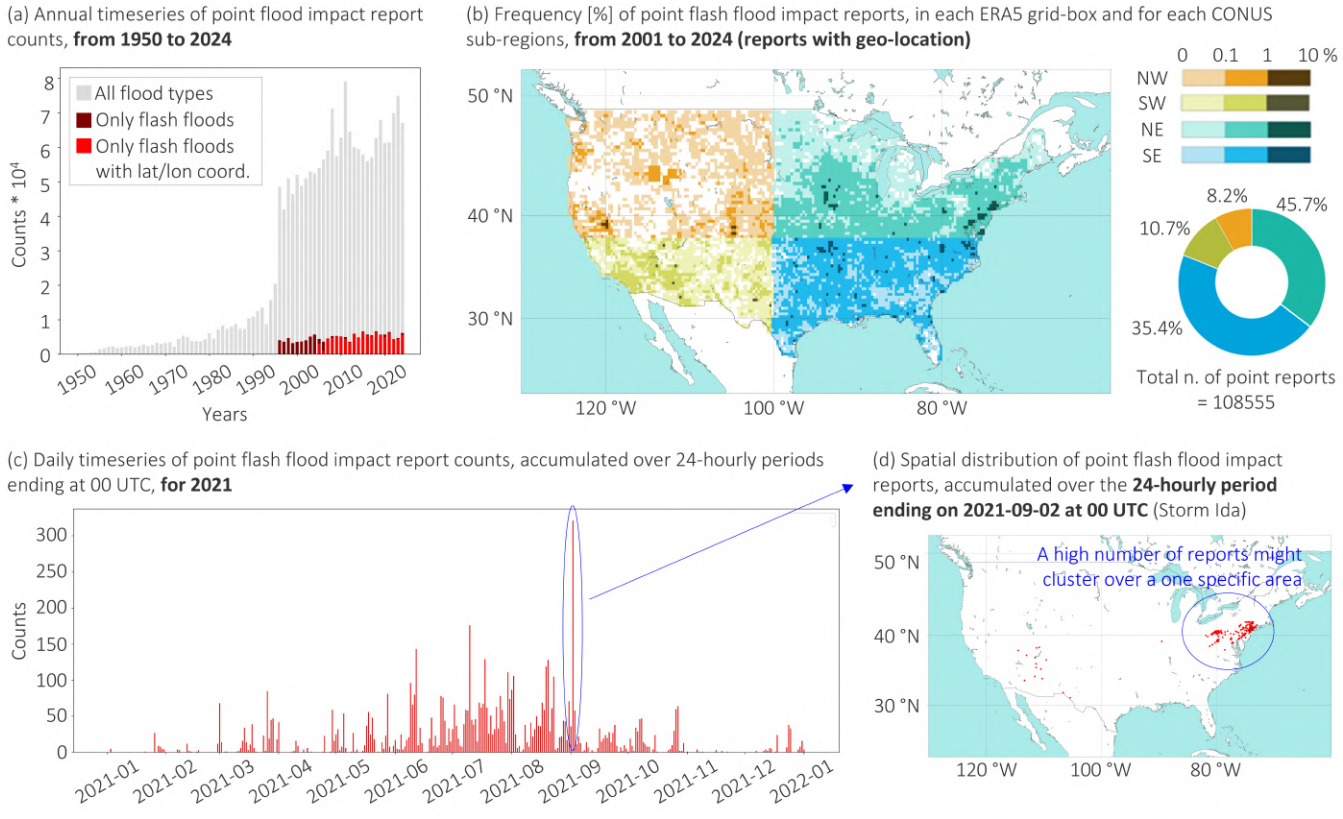

Figure 2. Point flash flood reports in the Storm Event Database over the CONUS. (a) Annual time series of point flood reports from 1950 to 2024: all flood types (grey bars), point flash flood reports (dark red bars, from 1996), and geo-located point flash flood reports (bright red bars, from 2001). (b) Frequency (%) of point flash flood reports per ERA5 grid-box and CONUS sub-region (NW in orange; SW in yellow; NE in green; SE in blue) from 2001 to 2024; the pie chart shows the regional distribution and total number of point reports (n=108,555). (c) Daily time series of point flash flood report counts (red bars), accumulated over 24-hourly periods ending at 00 UTC, for 2021; the blue circle marks the 24-hourly period ending on 2021-09-02 at 00 UTC (Storm Ida). (d) Spatial distribution of point reports (red dots) for the same day.

2.3 Model features

In numerous studies, four parameters consistently emerge as primary drivers of flash flood occurrence: extreme, localised rainfall (Schumacher, 2017), antecedent soil moisture (Grillakis et al., 2016), orography slope (Zhai et al., 2018), and vegetation coverage (Costache et al., 2020). These studies also show that rainfall acts as the principal trigger for flash flood occurrence, whilst the remaining three "non-meteorological" variables modulate the catchment response to flash-flood-producing rainfall events.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

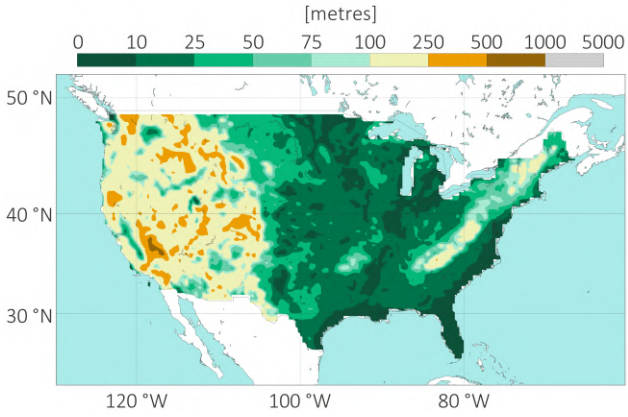

Figure 3. ERA5 standard deviation of the filtered sub-grid orography (SDFOR, in metres) — static field. Green colours indicate low terrain variability within ERA5 grid-boxes; warm colours (yellow/brown) indicate locally rugged relief; grey denotes values exceeding 1000 m.

2.3.1 Non-meteorological features

The non-meteorological features (antecedent soil moisture, orographic slope, and vegetation coverage represented by the leaf area index) are derived from ERA5 (Hersbach et al., 2020), which provides spatially complete and temporally consistent fields at 31 km resolution from 1940 to near real-time. ERA5's global coverage ensures that the predictors are available worldwide, should the model skill warrant global deployment. ERA5 *reanalysis* (from 2001 to 2020) was used for model training, and ERA5 *forecasts* (from 2021 to 2024, up to day 5) were used during inference. Technical details on the computation of these three features are provided in Appendix B, whilst their physical interpretation is described below, as it underpins the interpretation of the SHAP analysis (Section 4.1) and the case studies (Section 5).

Figure 3 shows the standard deviation of the filtered sub-grid orography (SDFOR) values over the CONUS. This parameter is called *static* as it does not changed in time. This parameter broadly mirrors the 1-km orography field in Figure 1a, with the highest values concentrated over the west side of the CONUS and a secondary maximum over the Appalachians in the east. The SDFOR parameter captures terrain variability within each ERA5 grid-box rather instead of absolute elevation. Hence, it may accentuate areas of locally rugged relief at lower levels, while high-elevation plateaux may have low SDFOR values, despite sitting at considerable altitude. For example, the highest values in the sub-grid orography field are not over the Rocky Mountains. They occur between the Death Valley and the Amargosa Range, in the Pacific coast of California, where elevations rise from approximately sea level to ∼2,600 m a.s.l. over a distance of less than 300 km, producing exceptionally high terrain variability within the ERA5 grid-boxes spanning this transition.

Figure 4 illustrates the spatial distribution of the percentage of maximum soil saturation (PMSS) across the CONUS, valid for 2021-09-01 at 00 UTC. Panel (a) presents the reference state derived from ERA5 reanalysis. Panels (b), (c), and (d) display ERA5 forecasts for the same valid time, but for different lead times: day 1 (t+24), day 3 (t+72), and day 5 (t+120). Heavy rainfall from Storm Ida, over the preceding days, caused a distinct area of near-total saturation (approaching 100%) in the

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

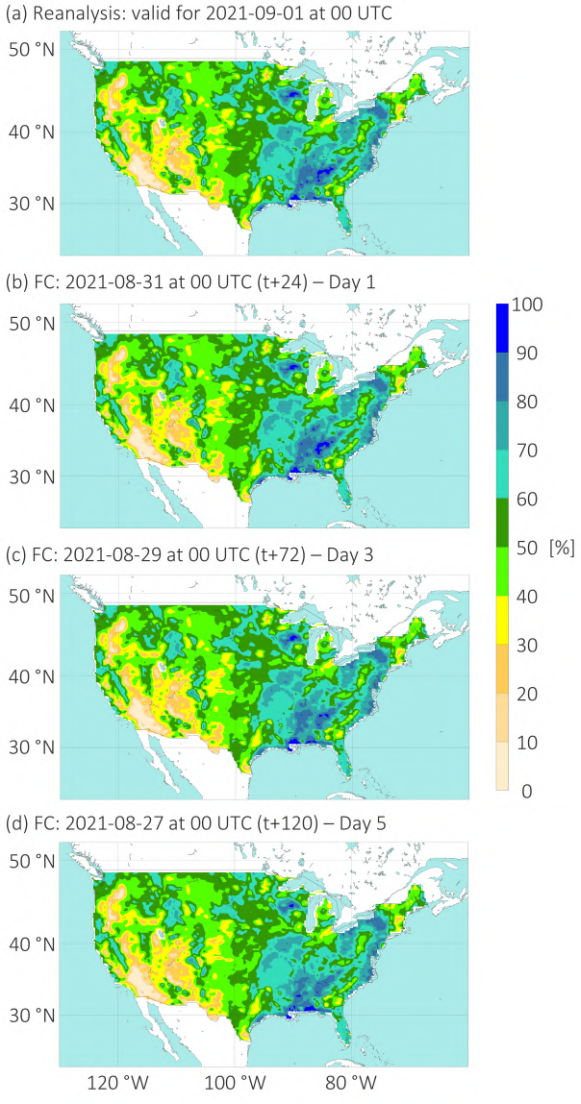

Figure 4. ERA5 percentage of maximum soil saturation (PMSS, in %), valid for 2021-09-01 at 00 UTC — dynamic field. (a) Reanalysis. (b)–(d) Forecasts (FC) for the same valid time, but initialised on (b) 2021-08-31 at 00 UTC (t+24, day 1), (c) 2021-08-29 at 00 UTC (t+72, day 3), and (d) 2021-08-27 at 00 UTC (t+120, day 5). Warm colours (brown/yellow) indicate dry conditions; cool colours (green/blue) indicate wetter conditions.

southeastern CONUS (Figure 4a). For this reason, PMSS is called *dynamic*: its spatial distribution evolves in response to antecedent hydro-meteorological conditions, capturing the progressive accumulation of soil moisture in the days preceding a potential flash flood event. The figure demonstrates the model's ability to capture this saturation pattern up to five days in advance (Figure 4b-d). Even though there is a gradual smoothing of the saturation pattern and a reduction in peak values, the location of the wettest regions in the forecast fields remains consistent with those in the reanalysis reference.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

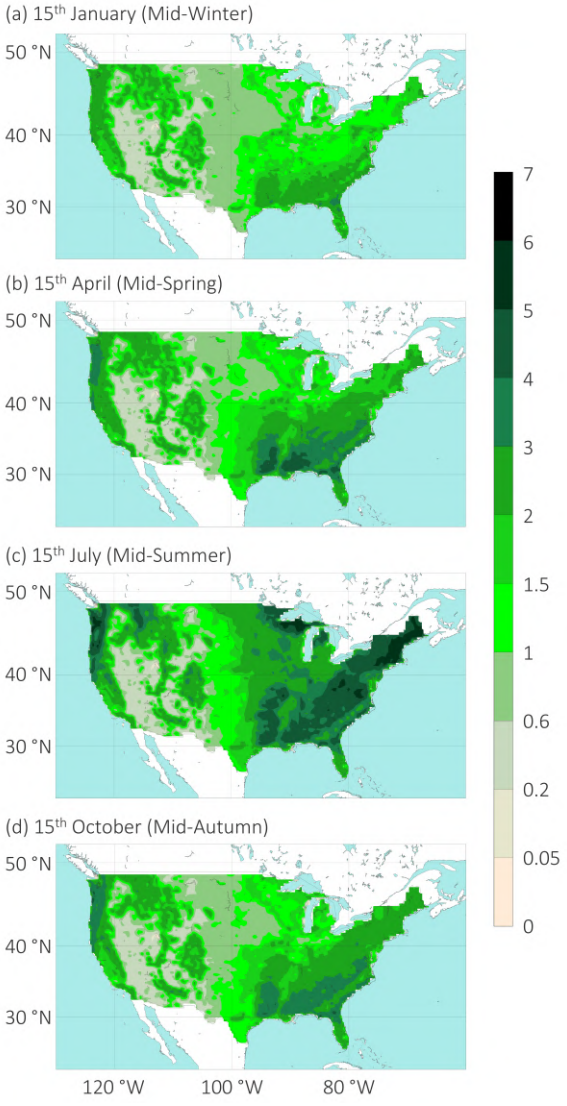

Figure 5. ERA5 leaf area index (LAI, dimensionless) — climatological field. (a) 15th January (mid-winter), (b) 15th April (mid-spring), (c) 15th July (mid-summer), and (d) 15th October (mid-autumn). Light beige indicates absent or sparse vegetation; bright green indicates moderate density; dark green indicates high density.

Figure 5 shows the spatial distribution of the leaf area index (LAI) over the CONUS. In ERA5, LAI values do not evolve in response to antecedent hydro-meteorological conditions. Lai is considered a *climatological* variable because its values are prescribed as a climatological variable, whose values vary only according to a fixed seasonal cycle derived from long-term averages. Hence, it remains identical across years for a given day of the year. In mid-winter (15th of January, Figure 5a), LAI is at its minimum. Most interior and northern regions show values near 0–0.2, indicating dormant vegetation or bare ground,

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

while evergreen forests in the Pacific North-West retain moderate coverage. By mid-spring (15th of April, Figure 5b), a green-up is evident in the South-East where values rise to ranges between 2 and 4. The cycle peaks in mid-summer (15th of July, Figure 5c), where vast areas of the eastern US and Pacific North-West reach saturation values ranging between 5 and 7. Finally, mid-autumn (15th of October, Figure 5d) marks the decline associated with leaf senescence, as values in northern latitudes range only between 1 and 2, although the South-East retains relatively high vegetation coverage.

2.3.2 Rainfall

Like any other model, the gridded rainfall dataset, ERA5, provides average rainfall estimates over the model grid-box. Due to its coarse spatial resolution (31 km), it smooths localised, high-intensity rainfall peaks that may trigger flash floods (Lavers et al., 2022). Moreover, ERA5's convection parametrisation Bechtold et al. (2014) introduces further biases, primarily due to the underrepresentation of the rainfall's diurnal cycle and the fact that convective cells remain anchored to their triggering grid column.

To better capture point-scale rainfall extremes, this study uses rainfall from ERA5-ecPoint, i.e., ERA5 rainfall estimates (from reanalysis and forecasts) post-processed with the ecPoint technique (Hewson and Pillosu, 2021). ecPoint is a statistical post-processing technique that transforms global gridded NWP outputs into probabilistic point-scale forecasts that mirror observations from rain gauges by addressing the two main factors affecting the performance of global NWP model outputs against point verification: systematic biases (Lavers et al., 2021) and lack of representation of sub-grid variability (Göber et al., 2008). Each raw, deterministic ERA5 realisation is converted into a distribution of N=100 point-scale rainfall estimates within each grid-box, and distilled into 99 percentiles, from 1st to 99th. Objective verification between raw ERA5 and ERA5-ecPoint reanalysis has shown that the latter represents light (0 mm) and extreme (i.e, rainfall exceeding the 10-year return period) point rainfall estimates better than ERA5 (Pillosu et al., 2025).

ERA5-ecPoint rainfall estimates (reanalysis and forecasts) are provided in the ERA5's native grid, i.e., reduced Gaussian grid N320, with a ∼31 km spatial resolution at the equator. This study considers 24-hourly rainfall accumulation periods ending at 00 UTC. The forecasts are produced up to day 5, with accumulations between (t+0,t+24) for day 1, (t+24,t+48) for day 2, (t+48,t+72) for day 3, (t+72,t+96) for day 4, and (t+96,t+120) for day 5.

Figure 6a and Figure 7a show the distribution over the CONUS (static maps) of point-scale rainfall estimates (in mm/24h) that correspond, respectively, to the 1- and 50-year return period (computed using the 1991-2020's climatology). Figure 6b and Figure 7a show the probability of ERA5-ecPoint rainfall reanalysis of exceeding, respectively, the 1- and 50-year return period, valid for the 24-hourly period ending on 2021-09-02 at 00 UTC. Figures 6c-e and Figures 7c-e show the probability forecasts, respectively, for day 1, day 3, and day 5. Two contrasting events illustrate how varies the predictability of different rainfall events. The north-east coast of CONUS was affected by Storm Ida, a large-scale convective system (highlighted by blue circles). For this event, exceedance probabilities for both return periods are uniformly elevated across an extensive area and remain robust (above 60%) at all lead times, indicating strong predictability for synoptically forced systems. By contrast, the south-west quadrant was affected by a storm-scale convective system (highlighted by red circles). Here, the probability of exceeding the 1-year return period averages roughly 20% at day 1 but falls to approximately 5% by day 5, and the location of

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

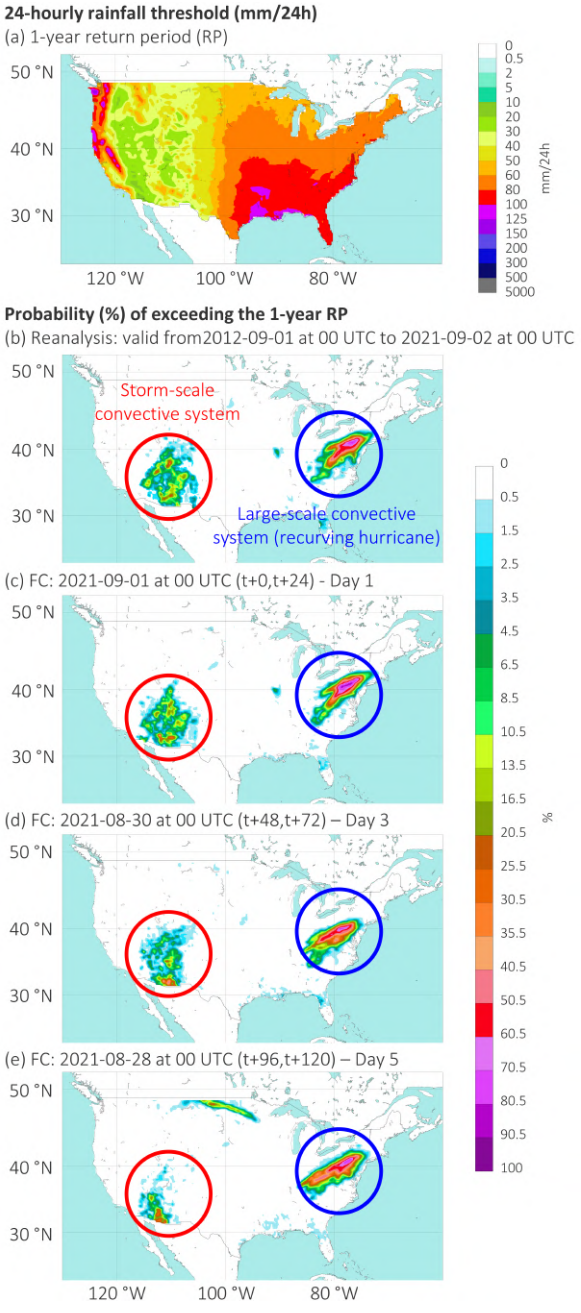

Figure 6. 24-hourly ERA5-ecPoint rainfall. (a) 1-year RP threshold (mm/24h) from ERA5-ecPoint reanalysis (1991–2020 climatology). (b) Probability (%) of exceeding the 1-year RP in reanalysis, valid for the 24-hourly period ending on 2021-09-02 at 00 UTC. (c)–(e) Same probabilities but for day 1 (FC: 2021-09-01 at 00 UTC (t+0,t+24)), day 3 (FC: 2021-08-30 at 00 UTC (t+48,t+72)), and day 5 (FC: 2021-08-28 at 00 UTC (t+96,t+120)) forecasts (FC).

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

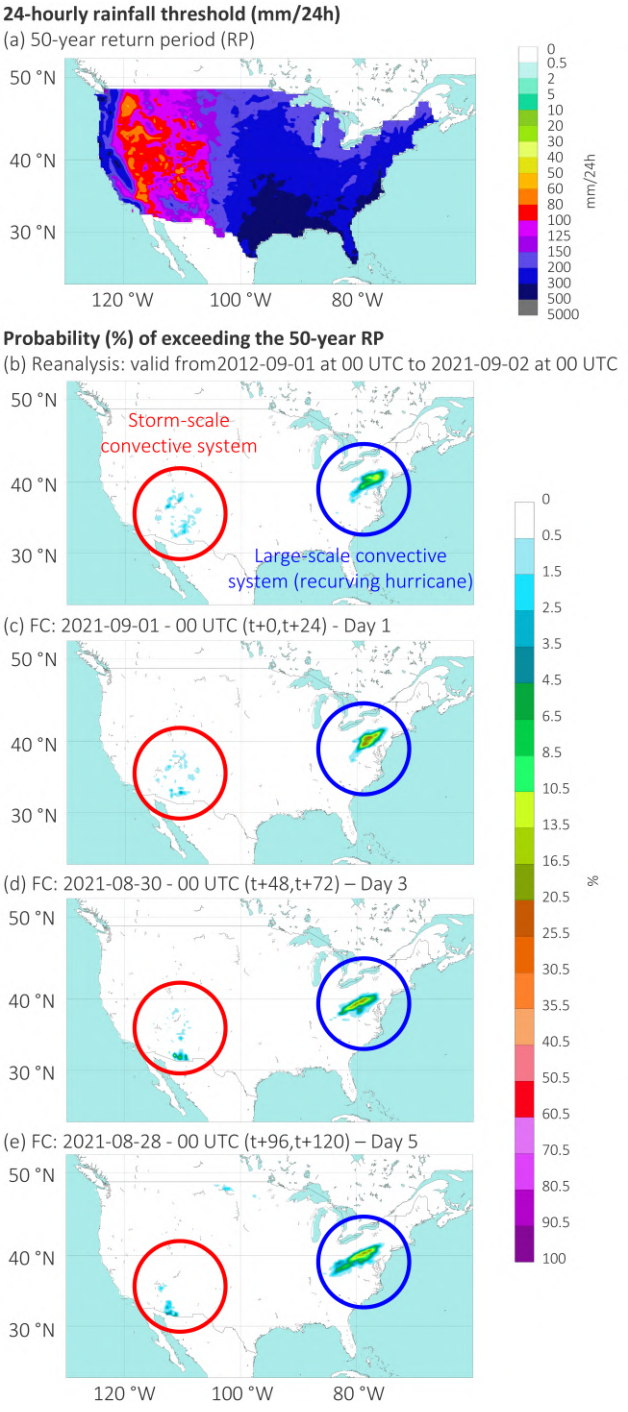

Figure 7. Same as Figure 6 but for the 50-year return period.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

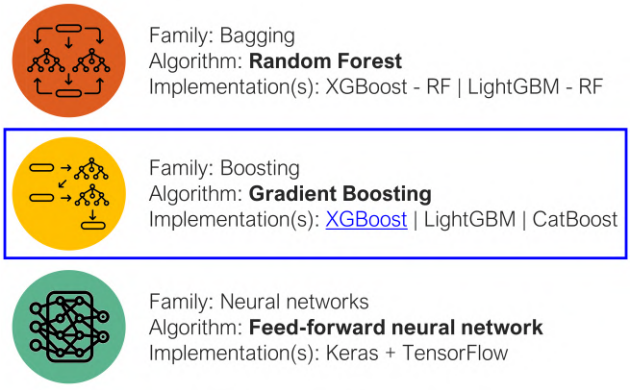

Figure 8. Overview of the three ensemble ML model families used in this study: bagging, via random forests (XGBoost and LightGBM implementations); boosting, via gradient boosting (XGBoost, LightGBM, and CatBoost implementations); and neural networks, via feed-forward architectures (implemented via Keras with TensorFlow backend). The best-performing family (boosting) is highlighted by the blue square; the best-performing implementation (XGBoost) is highlighted in blue.

the predicted extremes shifts considerably between successive forecast runs. A similar degradation is observed for the 50-year return period, where the spatial uncertainty in the predicted area of intense rainfall is even more pronounced.

3 Methods

3.1 Development of a ML model under an imbalanced training dataset

3.1.1 ML algorithm selection

Ensemble learning techniques manage class imbalance effectively (Ayodele, 2023; Altalhan et al., 2025). Within this category, three model families were considered (bagging, boosting, and neural networks, Figure 8), for their capacity to extract complex non-linear patterns from tabular inputs and generalise to unseen data (Shwartz-Ziv and Armon, 2022). Within a number of implementations for each model family (described in Appendix C1), XGBoost gradient boosting achieved the best overall performance in the prediction of flash flood occurrence (Pillosu, 2026). Hence, all subsequent analysis focuses exclusively on this implementation (highlighted by the blue square in Figure 8), whilst the full comparison across all evaluated architectures is provided in the Supplementary Material accompanying this manuscript.

3.1.2 Loss function selection

Two loss function configurations were evaluated to assess the sensitivity of predictions to class imbalance (Table C1 in Appendix C2): standard binary cross-entropy (BCE) and weighted binary cross-entropy (W-BCE), with the positive class weight treated as an optimisable hyperparameter (Table C2 in Appendix C2). An algorithmic approach, adjusting the loss function

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

directly, was preferred over synthetic data generation to avoid the additional uncertainty introduced by oversampling methods (Altalhan et al., 2025).

3.1.3 Training strategy: repeated stratified nested cross-validation

To address the severe class imbalance in the training dataset, repeated stratified nested cross-validation was performed using Scikit-Learn's "RepeatedStratifiedKFold"[6] (Figure 9). The nested structure — an outer loop for model generalisation assessment and an inner loop for hyperparameter tuning — prevents data leakage and overoptimistic performance estimates (Sasse et al., 2025; Ying, 2019). Stratification of all splits ensures that each fold preserves the class distribution of the original training dataset, avoiding dataset shift (López et al., 2014), which can distort performance measures under severe imbalance. The candidate model corresponds to the outer fold producing the highest performance on the held-out outer test subset, and is subsequently retrained on the full training dataset for operational deployment.

3.2 Hyperparameter tuning

Hyperparameter tuning was performed using the Python library called Optuna (Akiba et al., 2019). Unlike grid search (testing all combinations of hyperparameter values) or random search (testing random combinations), Optuna's Bayesian algorithm learns from previous trials to select which hyperparameter values to test next, navigating efficiently complex, continuous hyperparameter spaces even with limited computational resources. 20 distinct hyperparameter configurations were tested. The evaluated hyperparameters and their value spaces are shown in Table C2 in Appendix C3.

Two evaluation metrics were used to assess candidate hyperparameter configurations within the inner cross-validation loop: the area under the Receiver Operating Characteristic curve (AUC-ROC) and the area under the Precision-Recall curve (AUC-PR), whose formal definitions are provided in Appendix C4. The two metrics differ in their sensitivity to false alarms (i.e., when flash flood occurrence is predicted but it is not subsequently observed) under severe class imbalance. A model that maximises AUC-ROC tends to favour detection sensitivity (i.e., the model's ability to identify flash flood occurrences) even at the expense of a good reliability (i.e, the degree to which forecast probabilities correspond to observed occurrence frequencies). On the contrary, a model that maximises AUC-PR tends to favour reliability. Both metrics are carried forward into the objective verification to expose this trade-off explicitly.

3.3 Feature engineering

Feature engineering was minimal, with models trained on primarily raw variables except for one derived feature: the maximum probability of 24-hour rainfall exceeding the 1- and 50-year return period in adjacent grid-boxes. This feature addresses two structural limitations. First, convective parametrisation schemes in global NWP models generate stationary rainfall cells anchored to their triggering grid-column rather than propagating with the wind (Doswell, 2001), causing systematic underestimation of flash flood likelihood at the downwind coastline where convective systems deliver their rainfall. Second, the absence

[6] scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

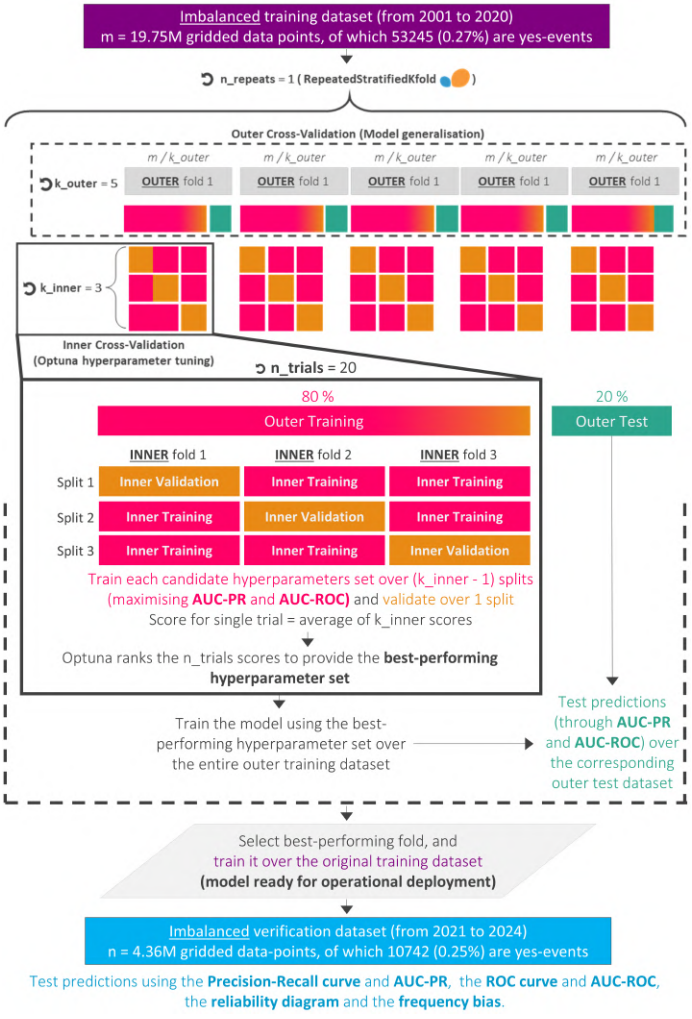

Figure 9. Workflow for a repeated stratified nested cross-validation. The whole dataset considered in this study (2001-2024) is divided into a **training dataset** (2001-2020) and an independent **verification dataset** (2021-2024). The workflow consists of an outer cross-validation loop to assess model generalisation and an inner cross-validation loop for hyperparameter tuning. The outer loop uses Scikit-Learn's RepeatedStratifiedKFold to create k_outer=5 **outer folds** across n_repeats=1 iterations. Each of them preserves the class distribution of the **training dataset**. Each **outer fold** is divided into an **outer test subset** (20%) and an outer training subset (80%). In the latter, a Bayesian hyperparameter tuning is performed via Optuna through inner cross-validation over n_trial=20 repetitions: each trial trains on **inner training folds** and validates on **inner validation folds**, with performance measured as the mean of AUC-ROC and AUC-PR over the **outer test subset**. The optimal hyperparameter set is used to train the final model on the complete outer training subset. Performance is assessed on the held-out **outer test fold** using AUC-ROC and AUC-PR. The best-performing fold is retrained on the full **training dataset** for operational deployment, and it is verified over the independent **verification dataset**, considering PR curves and AUC-PR, ROC curves and AUC-ROC, reliability diagrams, and frequency bias (FB).

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

of hydrological routing in this model, between grid-boxes, means that rainfall in one grid-box can cause flooding in adjacent downstream grid-boxes. Without this feature, flash flood occurrence in catchments of 100-500 km^2, where routing plays a critical role, would be systematically underestimated, as seen in the Valencia floods in 2024.

3.4 Objective verification of the flash flood occurrence forecasts

The verification of flash flood occurrence predictions focused on two desirable properties of probabilistic forecasts: reliability and discrimination ability (Jolliffe and Stephenson, 2012; Wilks, 2020). Each property was assessed at two levels of detail. Overall scores summarise performance across the full forecast probability distribution, enabling rapid model comparison. Breakdown scores decompose performance across individual probability thresholds instead, exposing probability-threshold-dependent behaviours that aggregated metrics may obscure.

Reliability was assessed using frequency bias (FB) as an overall score and reliability diagrams as breakdown scores. Discrimination ability was assessed using the Relative Operating Characteristic (ROC) curves as breakdown scores and the area under the ROC curve (AUC-ROC) as an overall score. To complement ROC-based scores when assessing forecasts computed with imbalanced training datasets, Precision-Recall (PR) curves are used as breakdown scores and the area under the PR curve (AUC-PR) as an overall score (Saito and Rehmsmeier, 2015). A detailed description of these scores is provided in Appendix C4.

Confidence intervals (99%) were computed via bootstrapping (1000 replicates) but were negligibly small and are omitted from subsequent figures.

3.5 Physical interpretation of the ML model outputs: SHAP

SHapley Additive exPlanations (SHAP) were used to physically interpret the data-driven model outputs. Derived from cooperative game theory, SHAP decomposes each prediction into additive contributions from individual features (Strumbelj and Kononenko, 2010; Rozemberczki et al., 2022), and is particularly well suited to tree-based models (Zhao et al., 2025c). In this study, SHAP values are used to assess which hydro-meteorological predictors dominate the prediction of flash flood occurrence, how features interact to modulate predicted probabilities, and whether the learned relationships are physically consistent with established hydro-meteorological understanding or reflect spurious correlations in the training data.

The SHAP analysis was conducted using flash flood occurrence probabilities generated by the best-performing model configuration identified in Section 3.1.1 (XGBoost with a loss function optimised for class imbalance and AUC-ROC as the evaluation metric) and forced exclusively with reanalysis data.

3.6 Global extension of the regional training

Three training approaches (TA) were tested to determine the optimal strategy to extend the regional training and develop predictions of flash flood occurrence across a continuous global domain.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

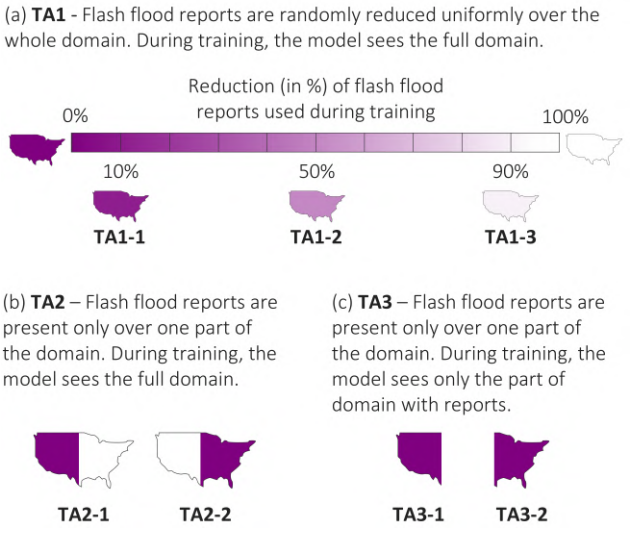

Figure 10. Training approaches (TA) for the sensitivity analysis on the global extension of regional training. (a) TA1: flash flood impact reports are reduced uniformly over the full domain by 10% (TA1-1), 50% (TA1-2), and 90% (TA1-3); the model sees the full domain during training. (b) TA2: flash flood impact reports are present only over one sub-domain (TA2-1, west; TA2-2, east), but the model sees the full domain during training. (c) TA3: flash flood reports are present only over one sub-domain (TA3-1, west; TA3-2, east), and the model sees only that sub-domain during training. The model training for each TA is carried over the **training dataset** (2001-2020).

The first approach (TA1) simulates training a global model over the full domain with flash flood reports available everywhere but at reduced density. This configuration reproduces the real scenario of training a global domain over the full domain using impact databases such as EM-DAT or Desinventar, which provide worldwide coverage but with considerably lower reporting density than regionally curated datasets such as NOAA's Storm Events Database. To assess whether training remains effective under different observational densities, flash flood reports were randomly sampled at three levels across the full CONUS domain, retaining 90% (TA1-1), 50% (TA1-2), and 90% (TA1-3) of the original dataset, whilst the model continued to train over the complete domain at each level (Figure 10a).

The second approach (TA2) simulates the scenario of training a global model over the whole global domain but with impact reports available only in certain regions. This configuration reproduces the real scenario of training a global model over the full domain using regional databases such as the NOAA Storm Events Database, which provides dense and reliable reports over the CONUS but contributes no flash flood labels over the rest of the world. Since the model cannot distinguish between regions where flash floods do not occur and regions where they occur but are unreported, absent labels are treated as true non-events during training. Such label absence introduces a systematic bias whereby, irrespective of the prevailing meteorological conditions, the model is penalised during training for predicting flash floods in unlabelled regions. That penalisation may suppress predictions even where learned hydro-meteorological relationships would otherwise indicate flash flood occurrence. The CONUS was divided into two configurations, a western region (west of 100°W, TA2-1) and an eastern region (east of

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

100°W, TA2-2), with flash flood observations restricted to one region whilst the model was trained over the full domain, exposing it to meteorological predictors from areas that contributed no positive labels during training (Figure 10b).

The third (TA3) approach simulates training a model exclusively over a data-rich region and deploying it over domains where no flash flood observations were available during training. This configuration reproduces the real scenario of training a regional model on a well-monitored area such as the CONUS, and subsequently deploying it globally, where flash flood reports are absent or insufficient for independent model development. Unlike the second approach, where the model is exposed to the full domain during training but receives no positive labels outside the reporting region, here the model sees only the sub-domain for which reports exist. The CONUS was again divided into two configurations, the western (TA3-1) and eastern (TA3-2) regions, with the model trained on flash flood observations from one region and subsequently applied to predict flash flood occurrence across the full domain, including areas unseen during training (Figure 10c).

As for the SHAP analysis, this sensitivity analysis was conducted using the best-performing model configuration forced with reanalysis data. The objective verification is presented in Section 4.3. Section 5.1 complements the objective verification analysis by comparing the probabilities of flash flood occurrence computed with the different training approaches to those computed with the full training dataset over the CONUS domain. Given the absence of sufficiently dense observational data for robust objective verification over the global domain, the performance of the medium-range forecasts is evaluated only for the Valencia flash floods on October 2024 (Section 5.2). Only the optimal training approach will be considered in the case study.

4 Results

4.1 Regional training: performance of probabilities of flash flood occurrence for reanalysis and medium-range forecasts

Six model implementations (i.e., XGBoost, LightGBM, and CatBoost for the gradient boosting algorithm, XGBoost and LightGBM for the random forest algorithm, and a feed-forward neural network) were evaluated using nested cross-validation with both balanced and weighted loss functions (full comparative results are provided in the accompanying Supplemental Material). Among these, the XGBoost implementation of gradient boosting trained with a balanced loss function and optimised for AUC-ROC achieved the best performance in terms of discrimination ability and reliability, whilst maintaining the shortest training times per fold. The remainder of this section presents the training diagnostics, independent verification, sensitivity analyses for global expansion of regional training, and case studies only for the selected XGBoost configuration.

During the nested cross-validation, the hyperparameter importance analysis (Figure 11a) reveals that maximum tree depth and learning rate are the two most influential hyperparameters for XGBoost, with normalised absolute Pearson's r coefficients of approximately 0.3, substantially exceeding those of the sub-sample ratio, column sampling rate, and number of estimators (all below 0.2). This indicates that model performance is governed primarily by the complexity of individual trees and the step size of the boosting procedure, whilst the stochastic regularisation parameters and ensemble size exert comparatively limited influence. Model generalisation (Figure 11b) was stable across all five outer folds, with AUC-ROC values consistently between approximately 0.78 (on average for the inner validation folds, line in orange) and 0.82 (for the outer test dataset, line

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

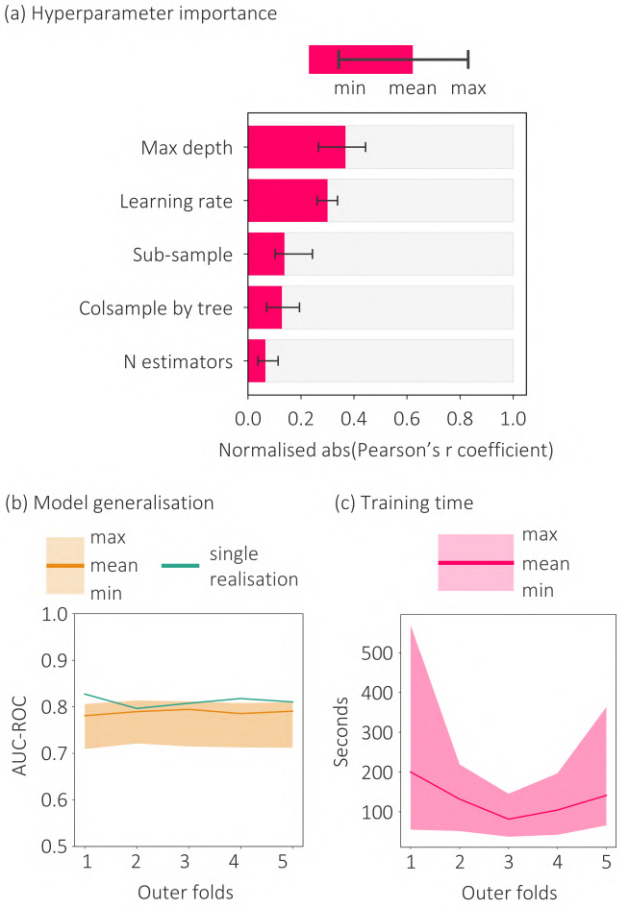

Figure 11. Training outcomes from repeated stratified nested cross-validation for XGBoost gradient boosting trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. (a) Hyperparameter importance over the **inner training folds**. (b) Model generalisation over the **inner validation folds** and the **outer test dataset**. (c) Training times over the **inner training folds**.

in green). The narrow range between the minimum and maximum realisations, and the close agreement between the AUC-ROC computed over the outer test dataset and the fold-mean, confirms that the nested cross-validation procedure produces robust performance estimates that are not unduly sensitive to the particular train–test split. The training time (Figure 11c) varied across the five outer folds, with mean values ranging from approximately 150 to 500 seconds per fold. The spread between minimum and maximum realisations was considerable for certain folds, reflecting sensitivity to the particular data partition, but it remained under 10 minutes. It is worth noting that amongst all the implementations, the feed-forward neural network achieved similar performance to the XGBoost implementation of gradient boosting here considered, but with a training time of approximately 8 hours per fold (see Supplemental Material). The best-performing outer fold was selected based on the combination of generalisation performance and computational cost. The selected fold (fold n. 4) comprises 343 shallow trees

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table 2. Optimal hyperparameters (fold n.4) for the XGBoost implementation of gradient boosting, trained with a balanced loss function and hyperparameters optimised for AUC-ROC.

Hyperparameter	Value
Maximum tree depth	4
Learning rate	0.029
Sub-sample ratio	0.980
Column sampling rate	0.986
Number of estimators	343

(maximum depth of 4) with a low learning rate of 0.029. The sub-sample ratio and column sampling rate are both close to unity, 0.980 and 0.986, respectively (Table 2).

The best-performing outer fold was retrained on the complete training dataset (2001-2020) and verified against the independent verification dataset (2021-2024). Flash flood occurrence maintains stable AUC-ROC around 0.8 (Figure 12a) and high AUC-PR values, from 0.03 to 0.01 (Figure 12b), when computed with reanalysis and forecast data, across all lead times. The frequency bias is around 1.2 for the flash flood occurrence computed with reanalysis data, and it is mostly stable at 2.5 for forecasts, at all lead times (Figure 12c). The ROC curve for day 1 forecasts (AROC = 0.803, Figure 12e) shows a similar behaviour to that for reanalysis data (AROC = 0.797, Figure 12d), and shows a fairly small deterioration over increasing lead times (AROC = 0.765 for day 3 (4% reduction compared to the day 1 forecasts) - Figure 12f - and AROC = 0.746 for day 5 (7% reduction)- Figure 12g). The precision-recall curve for day 1 forecasts (Figure 12i) also shows a very similar behaviour to that for reanalysis data (Figure 12h), except for very small values of recall, where in the precision-recall curve for reanalysis data the precision is 2.5 times bigger. As in the ROC curve, the precision-recall curves also show only a fairly small deterioration over increasing lead times (Figure 12j-k). As in the two previous scores, also the reliability diagram for day 1 forecasts (Figure 12m) shows a similar behaviour to that for reanalysis data (Figure 12l), with reliable forecasts for probabilities under 10%. For greater probabilities, the reliability diagrams tend to overestimate the observed probabilities of flash flood occurrence. For increasing lead times (day 3 in Figure 12n and day 5 in Figure 12o), such a threshold reduces to 2%.

4.2 Regional training: physical interpretation of the forecasts of probabilities of flash flood occurrence

The global feature importance ranking (Figure 13a) shows that the rainfall probability of exceeding the 1-year return period is the most important feature when estimating the probabilities of flash flood occurrence, contributing to 80% of the mean absolute SHAP values. Features regarding the vegetation cover (LAI, 35%), the orography steepness (SDFOR, 25%), the antecedent percentage of maximum soil saturation (PMSS), and the rainfall probabilities of exceeding the 1-year return period in adjacent grid-boxes (tp_prob_1_adj_gb, 12%) are also considered important by the model. The features related to the rain-

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

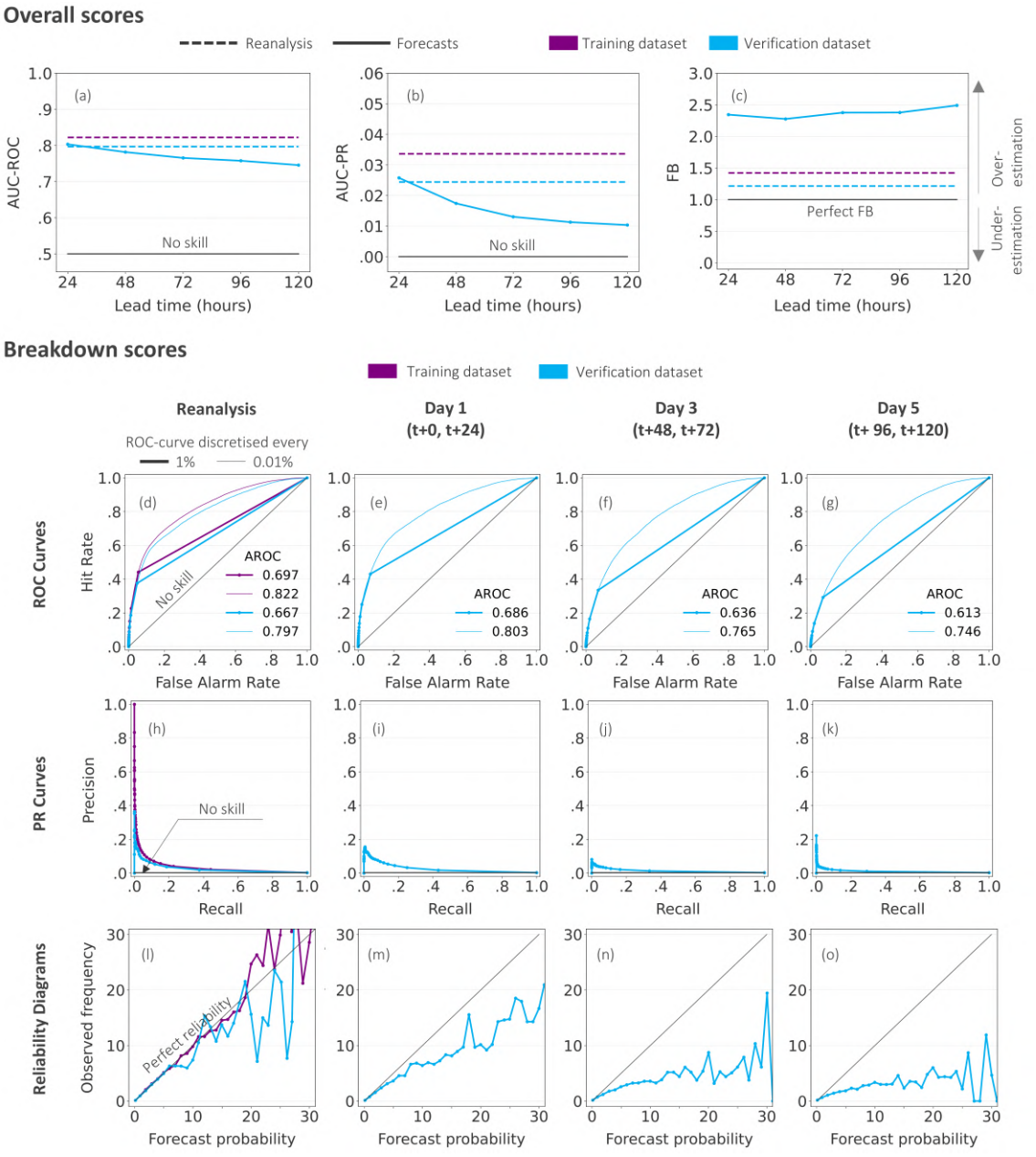

Figure 12. Objective verification. Flash flood occurrence computed using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. Reanalysis scores are shown for the **training dataset** and **verification dataset** with dashed lines. Continuous lines represents forecasts scores for the **verification dataset** only. Overall scores: (a) AUC-ROC (ROC curve discretised every 0.01%), (b) AUC-PR, and (c) FB. Breakdown scores for reanalysis (first column) and day 1, 3, and 5 forecasts (second to fourth columns): (d)–(g) ROC curves (thick line: 1% discretisation; thin line: 0.01%), (h)–(k) PR curves, and (l)–(o) reliability diagrams.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

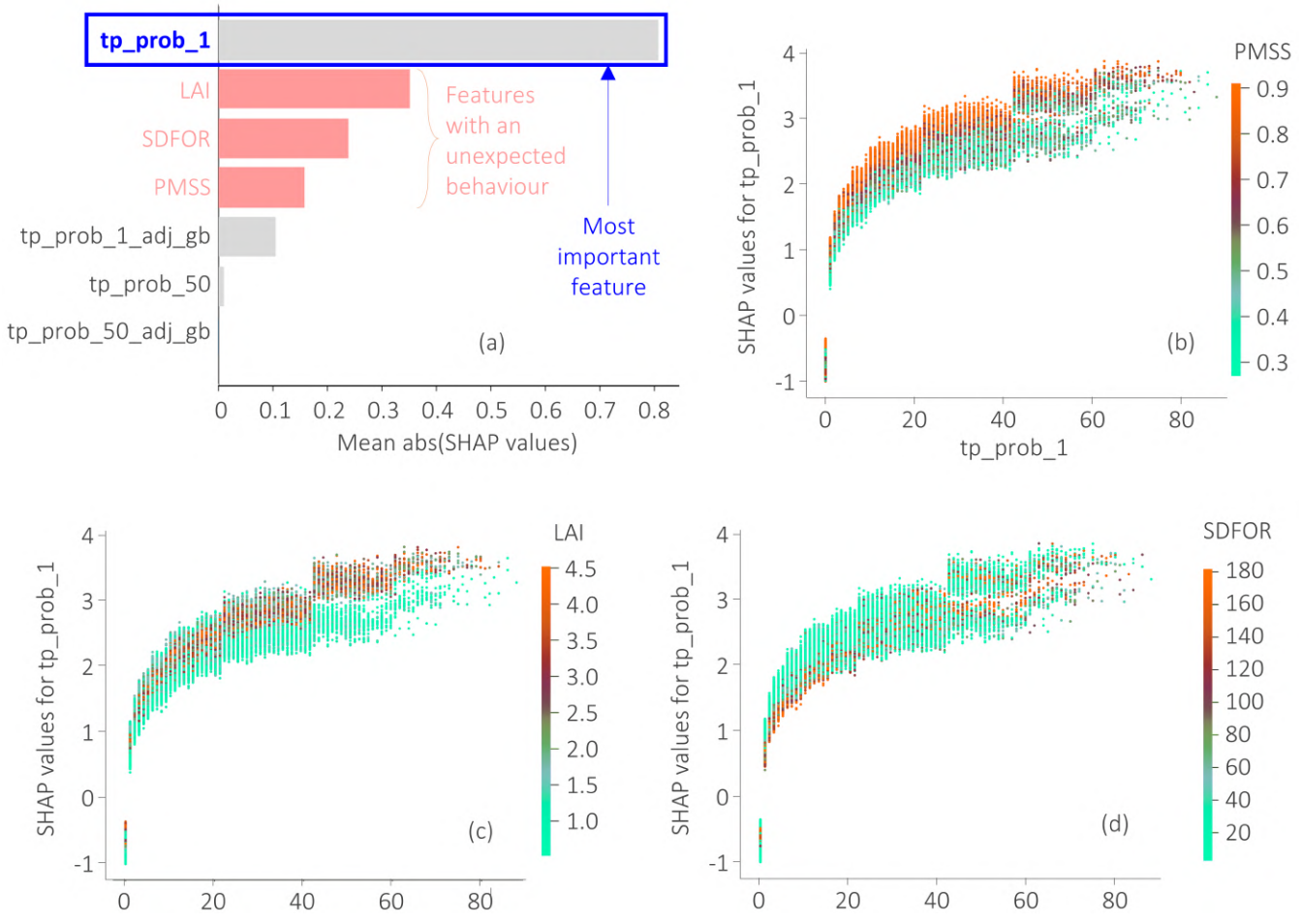

Figure 13. SHAP values. Flash flood occurrence computed for reanalysis data over the **verification dataset** (2021 - 2024), using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. (a) Global feature importance ranking (descending order); the most important feature is highlighted in blue, and features showing unexpected behaviour in pink. (b)–(d) Dependency plots between tp_prob_1 and, respectively, PMSS, LAI, and SDFOR.

fall probabilities of exceeding the 50-year return period (tp_prob_50 and tp_prob_50_adj_gb) are considered, overall, least important by the model to identify areas at risk of flash flood.

The dependency plots show critical threshold behaviours in the model's decision-making process. Overall, the rainfall-related feature tp_prob_1 (Figure 13b-d) contributes positively to the probabilities of having a flash flood in a grid-box. This contribution increases rapidly (from 0 to +3%) from probabilities between 0% and ~20%. For greater probabilities (>20%), the contribution of the rainfall parameters plateaus. The interactions of the rainfall-related variables with other features (expressed through the colour gradients) demonstrate that environmental conditions modify the contributions of rainfall in predicting areas at risk of flash floods. Areas with dense vegetation coverage (higher LAI values, Figures 13b and e), primarily flatter

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

(lower SDFOR values, Figures 13c and f), and with mostly saturated soils (higher PMSS values, Figures 13d and g) enhance sensitivity to rainfall, meaning that lower values of tp_prob_1 are required to trigger higher (positive) SHAP contributions in the probability of having a flash flood in a grid-box.

4.3 Global extension of the regional training

The AUC-ROC (Figure 14a, continuous line) and AUC-PR (Figure 14b, continuous line) show that TA1 and TA3 retain discrimination ability close to that of the full-dataset model (dashed lines), whereas TA2 degrades performance most substantially across all variants. For the FB (Figure 14c, continuous line), the three approaches exhibit markedly different behaviours. TA1 shows a progressive reduction in FB with increasing reductions of flash flood reports, whilst the spatial origin of the retained reports determines the direction of the FB in TA2 and TA3: variants trained exclusively on western CONUS reports (TA2-1, TA3-1) exhibit a stronger systematic underestimation, whereas those trained on eastern CONUS reports (TA2-2, TA3-2) produce FB values closest to the full-dataset reference (Figure 14c, dashed line). The single worst-performing approach for all three overall scores is TA2-1 (highlighted by the pink rectangle in Figures 14a-c), which retains only reports from the western CONUS but exposes the full domain during training. It reduces AUC-ROC by approximately 29% relative to the full-dataset model (from 0.8 to 0.57), collapses AUC-PR by roughly 90% (from 0.025 to 0.0024), and produces an FB of 0.0022. Conversely, TA3-2 (which retains only eastern CONUS reports and exposes only the grid-boxes in that domain during training, and highlighted by the brown rectangle in Figures 14a-c) achieves the closest overall performance to the full-dataset baseline across all three scores.

The breakdown verification scores (Figure 15) reveal how each training approach changes the discrimination ability and the reliability across probability thresholds. For TA1 (Figures 15a–c), the ROC curves built with 0.01% discretisation (thinner lines) remain close to that of the reference model trained on the full training dataset (Figure 12d, in cyan); however, when probability thresholds are discretised every 1% (thicker lines), the curves compress towards the bottom-left corner of the unit square, reducing AUC-ROC from 0.652 for TA1-1 to values approaching the no-skill threshold for TA1-2 (0.588) and TA1-3 (0.512). Among the spatially partitioned approaches, TA2-1 (Figure 15d) and TA3-1 (Figure 15f), both trained on western CONUS reports, produce the poorest ROC curves, with TA2-1 showing essentially no discrimination ability. In contrast, TA2-2 (Figure 15e) and TA3-2 (Figure 15g) yield the closest ROC curves to the reference model trained on the full training dataset (Figure 12d, in cyan). The PR curves reinforce this pattern. For TA1 (Figures 15h–j), reducing flash flood impact reports raises precision at very small recall values but progressively shrinks the achievable range of recall, with a recall value close to 0 for TA1-3 (Figure 15j). TA2-1 (Figure 15k) and TA3-1 (Figure 15m) achieve small values of recall, with TA2-1 also producing near-zero precision. TA2-2 (Figure 15l) and TA3-2 (Figure 15n) attain similar recall to the reference (Figure 12h, in cyan), but achieve greater values of precision. The reliability diagrams show that TA1-1 (Figure 15o) and TA1-2 (Figure 15p) maintain a comparable reliability to the full-dataset model (Figure 12l, in cyan), but with reliable probabilities below 20% and 15%, respectively. TA1-3 (Figure 15q) produces no reliable probabilities. TA2-1 (Figure 15r) produces an almost non-existent reliability diagram, confirming that this approach yields negligible probabilities of flash flood occurrence, whilst TA3-1 (Figure 15t) is able to produce a reliability diagram with probabilities that remain reliable below 10%. TA2-2 (Figure 15s) and TA3-2

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

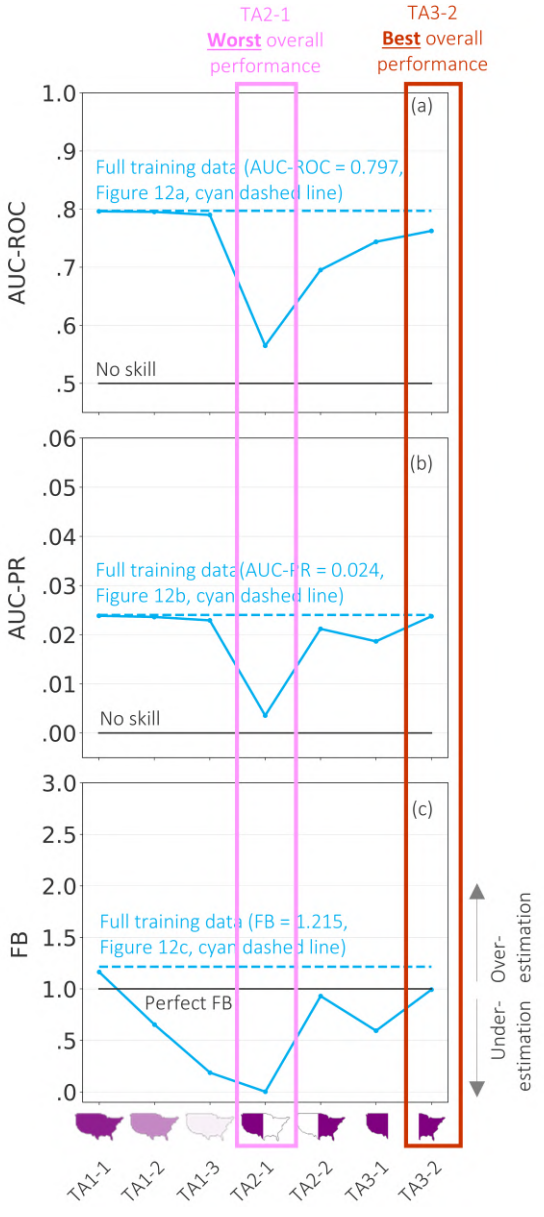

Figure 14. Overall verification scores for the sensitivity analysis on the global extension of regional training. Flash flood occurrence computed for reanalysis data over the **verification dataset** (2021 - 2024), using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. (a) AUC-ROC, (b) AUC-PR, and (c) (FB); solid lines show training approaches TA1–TA3, dash-dot lines show the reference for the model trained on the full training dataset (cyan dashed lines in Figures 12a-c). The pink and brown rectangles highlight, respectively, the worst (TA2-1) and best (TA3-2) performing approaches.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

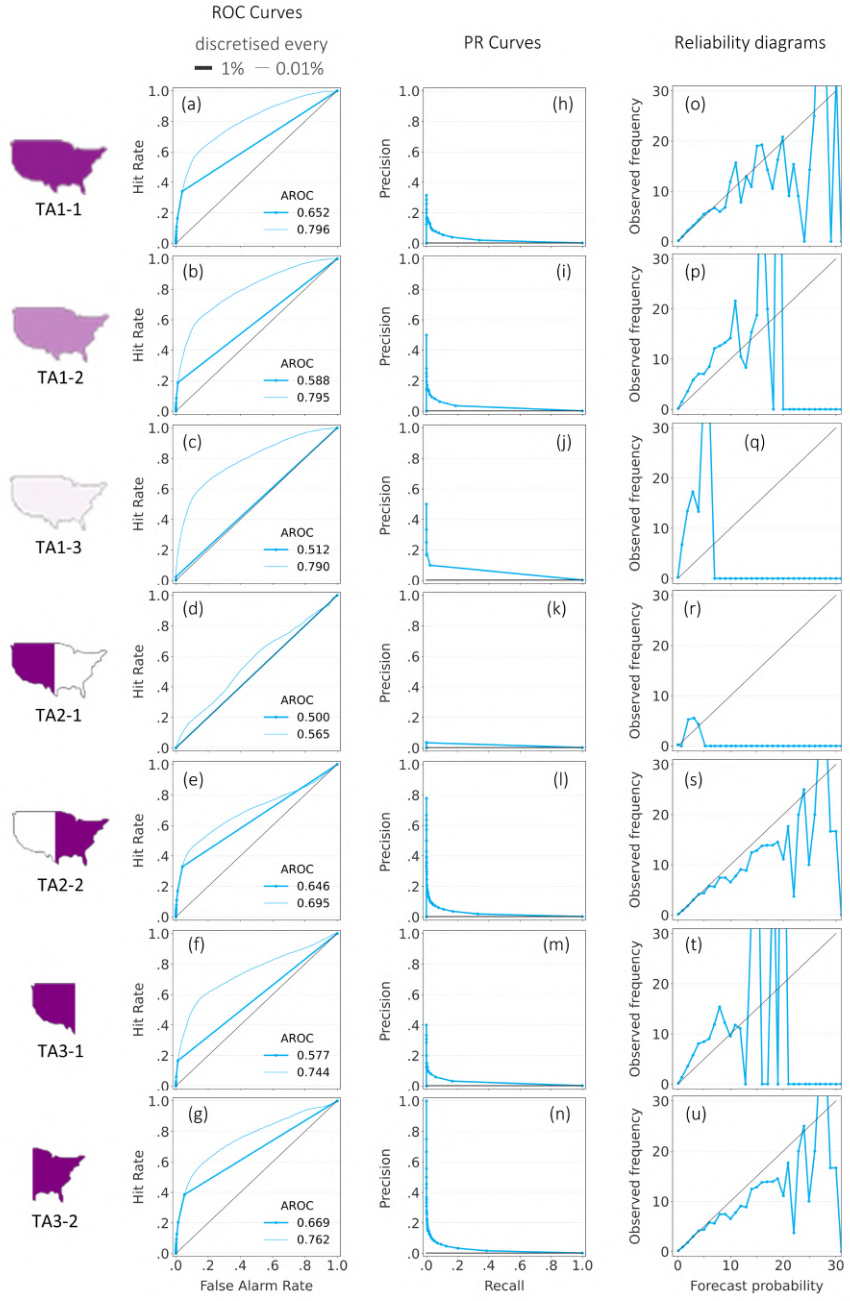

Figure 15. Breakdown verification scores for the sensitivity analysis on global extension of regional training. Flash flood occurrence computed for reanalysis data over the **verification dataset**. (a)-(g) ROC curves (thick line: 1% discretisation; thin line: 0.01%), (h)-(n) PR curves, and (o)-(u) reliability diagrams for training approaches TA1 (first to third rows), TA2 (fourth and fifth rows), and TA3 (sixth and seventh rows).

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

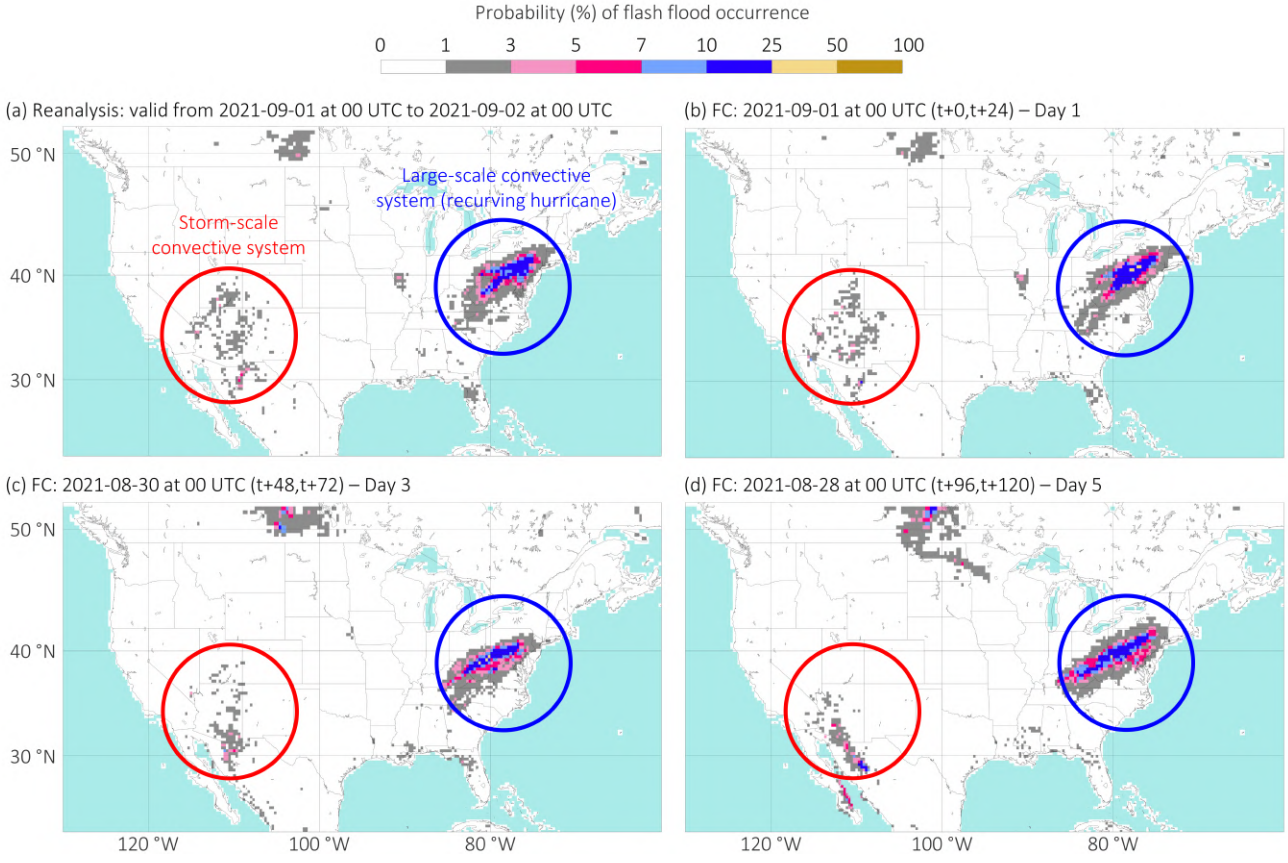

Figure 16. Probability (%) of flash flood occurrence computed using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. Flash flood occurrence is computed using the whole **training dataset**. All panels are valid for the 24-hourly period from 2021-09-01 at 00 UTC to 2021-09-02 at 00 UTC (Storm Ida). (a) Reanalysis. (b)-(d) Forecasts (FC): (b) 2021-09-01 at 00 UTC (t+0,t+24) - day 1, (c) 2021-08-30 at 00 UTC (t+48,t+72) - day 3, and (d) 2021-08-28 at 00 UTC (t+96,t+120) - day 5.

(Figure 15u) produce the closest reliability diagrams to the reference (Figure 12l, in cyan), with reliable probabilities up to 20%.

5 Case studies

5.1 Regional training and sensitivity analysis for global expansion of the regional training: Storm Ida

On 1 September 2021, the remnants of Hurricane Ida made landfall as a post-tropical cyclone over the north-eastern United States, producing extreme rainfall and widespread flash flooding from the Gulf Coast to New York. The Storm Event Database recorded over 200 flash flood reports for the 24-hourly period ending on 2021-09-02 at 00 UTC (Figure 2d), making it the single most reported flash flood day in 2021 (Figure 2c). Figure 16 shows the probabilities of flash flood occurrence for the

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

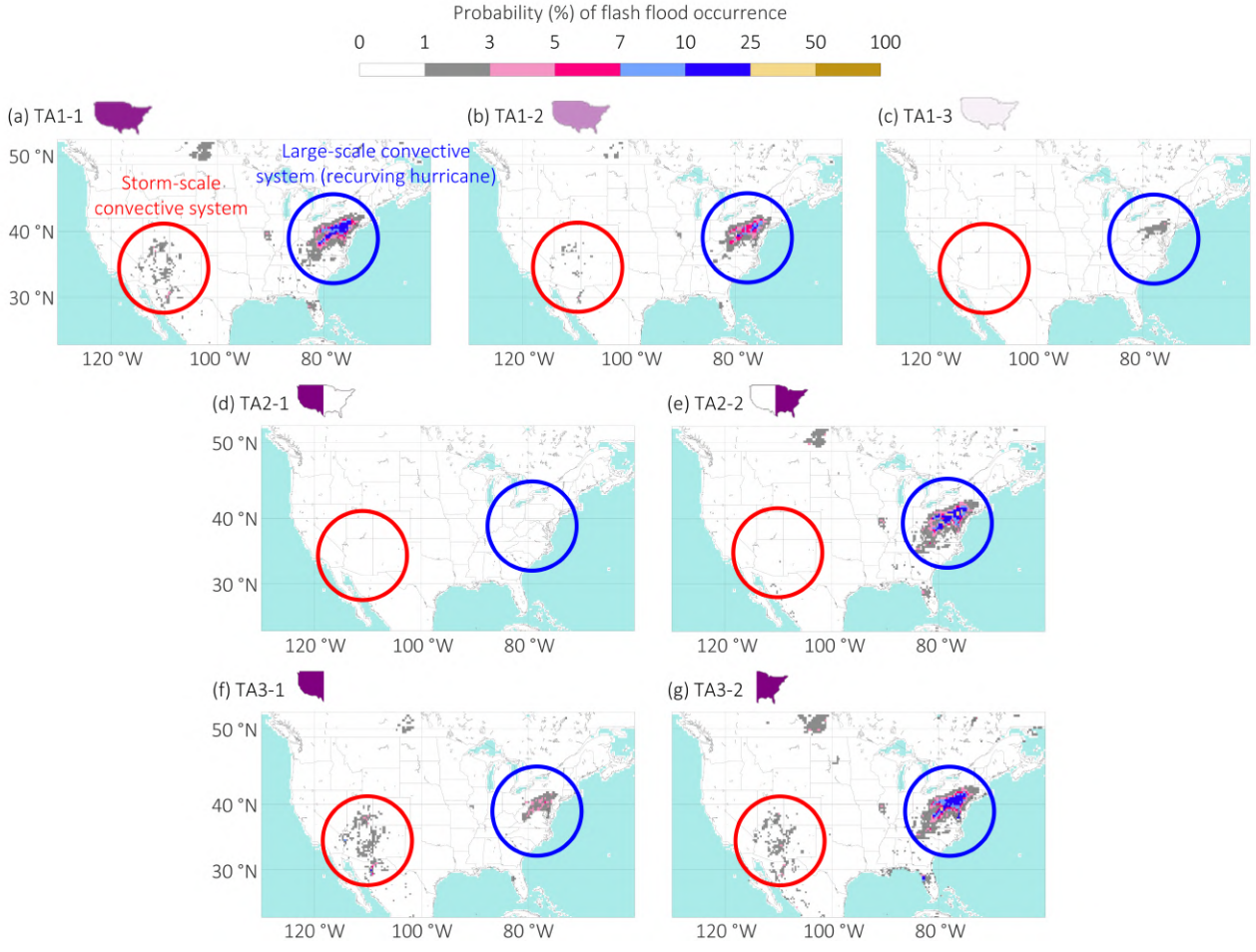

Figure 17. Probability (%) of flash flood occurrence for different training approaches. Flash flood occurrence computed for reanalysis data, using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. Valid for the 24-hourly period from 2021-09-01 at 00 UTC to 2021-09-02 at 00 UTC (Storm Ida). (a)–(c) TA1-1, TA1-2, and TA1-3. (d)–(e) TA2-1 and TA2-2. (f)–(g) TA3-1 and TA3-2.

same 24-hourly period. In all panels, two distinct signals are visible: a coherent band of high probabilities of flash flood occurrence (>10%) extending from the Gulf Coast north-eastward through New York (blue circle), produced by the large-scale convective system associated with Storm Ida, and a more diffuse, scattered signal (1–3%) over the central-western CONUS (red circle), associated with smaller-scale convective activity. The reanalysis-driven estimates (Figure 16a) correctly identify the areas where flash floods were reported in the Storm Event Database (Figure 2d). Comparing these probabilities of flash flood occurrence with the underlying ERA5-ecPoint rainfall inputs provides insight into the added value of the data-driven model. The probability of rainfall exceeding the 1-year return period (Figure 6b) is high over both events: exceeding 50% over much of the Storm Ida track (blue circle) and ranging between 3% and 10%, with localised peaks around 50%, over the

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

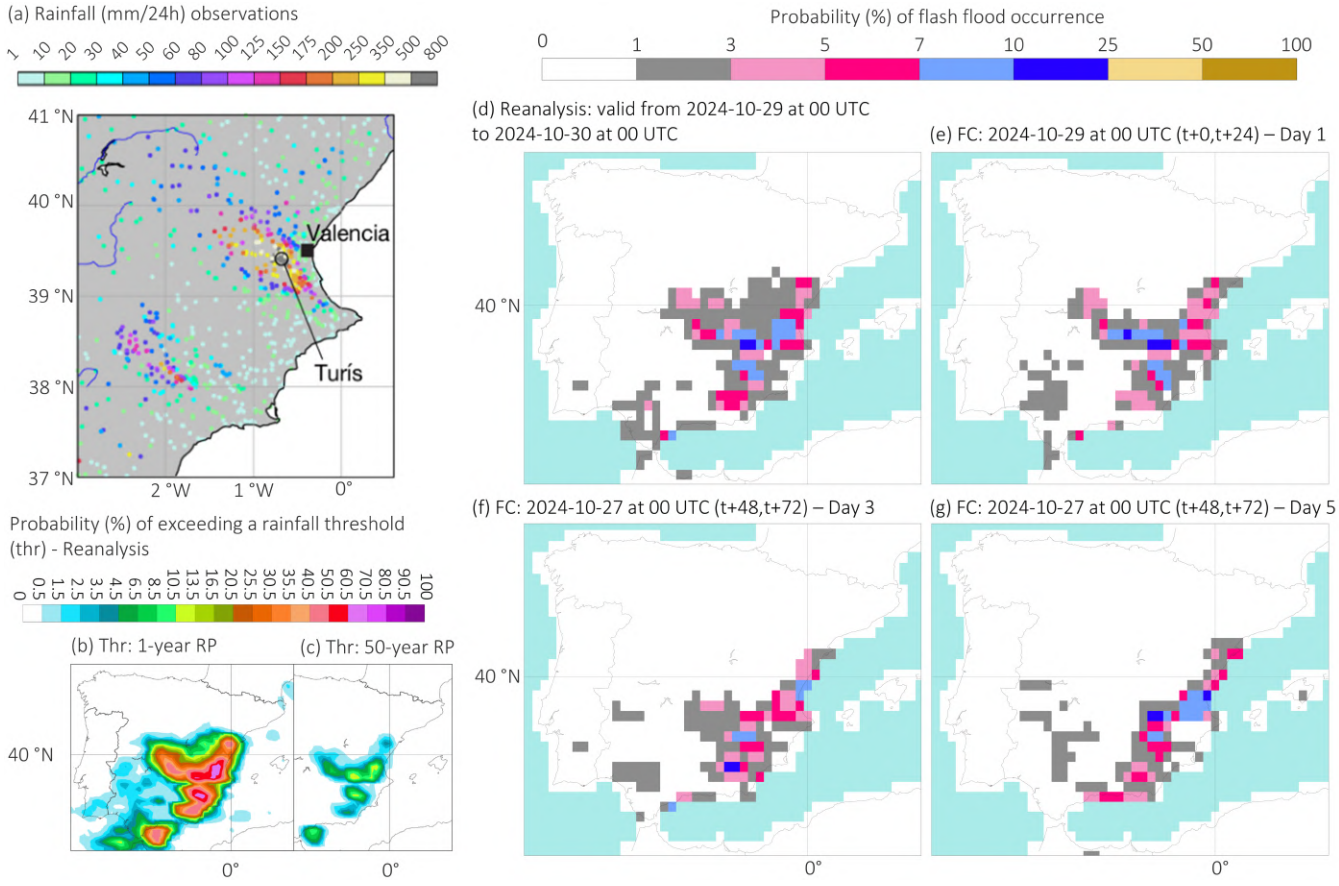

Figure 18. Probability (%) of flash flood occurrence in Valencia, Spain, using the XGBoost model trained with balanced-data loss function (BCE) and hyperparameters optimised for AUC-ROC. All panels are valid for the 24-hourly period from 2024-10-29 at 00 UTC to 2024-10-30 at 00 UTC. (a) Observed rainfall (mm/24h) from Gascón et al. (2025) (zoomed over Valencia). (b) Probability (%) of flash flood occurrence from reanalysis data (zoomed over Spain). (c)–(e) Probability of flash flood occurrence from forecasts (FC): (c) 2024-10-29 at 00 UTC (t+0,t+24) - day 1, (d) 2024-10-27 at 00 UTC (t+48,t+72) - day 3, and (e) 2024-10-27 at 00 UTC (t+96,t+120) - day 5.

storm-scale system (red circle). When considering the 50-year return period (Figure 7b), the two events diverge markedly: probabilities over the Storm Ida track remain around 10%, with peaks of 20–25%, whereas those over the western event drop to 0.5–1.5%. The flash flood occurrence model produces a sharper contrast between the two events than even the 50-year return period exceedance probabilities, with probabilities exceeding 25% over the Storm Ida track but remaining below 3% over the western system (Figure 16a). This amplification arises from the model's integration of non-meteorological features that modulate the catchment response to rainfall. The eastern CONUS, where Storm Ida struck, is characterised by relatively low sub-grid orographic variability (SDFOR predominantly below 50 m; Figure 3), near-saturated soils (PMSS exceeding 70% in the south-eastern quadrant; Figure 4a), and dense vegetation cover (LAI values of 5–7 in early September; Figure 5c). In contrast, the western CONUS exhibits rugged terrain (SDFOR exceeding 100–500 m; Figure 3), substantially drier antecedent

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

conditions (PMSS of 20–40%; Figure 4a), and sparser vegetation (LAI of 0.5–2; Figure 5c). The SHAP analysis (Section 4.2) confirms that these environmental contrasts are central to the model's predictions: areas with denser vegetation (higher LAI), flatter terrain (lower SDFOR), and wetter soils (higher PMSS) require lower rainfall exceedance probabilities to trigger positive SHAP contributions to flash flood occurrence (Figures 13b–d). The eastern CONUS thus provides exactly the combination of environmental conditions that the model has learned to associate with elevated flash flood risk, amplifying the rainfall signal relative to what would be inferred from rainfall exceedance probabilities alone. As noted in Section 4.2, the positive association between dense vegetation, flat terrain, and flash flood occurrence is physically counterintuitive — steep, sparsely vegetated catchments are traditionally considered more susceptible to flash flooding — and likely reflects the population-density bias in the Storm Event Database (Section 2.2), whereby the densely populated eastern CONUS contributes a disproportionate share of flash flood reports (Figure 2b). The day 1 forecast (Figure 16b) closely reproduces the reanalysis pattern, maintaining elevated probabilities over both the Storm Ida track (blue circle) and the storm-scale event (red circle). By day 3 (Figure 16c) and day 5 (Figure 16d), the large-scale signal remains identifiable although probabilities decrease progressively, whereas the signal associated with the smaller-scale convective activity becomes discernible only from day 3 (Figure 16c). This behaviour is consistent with the higher predictability of synoptically forced systems compared with storm-scale convection at medium-range lead times, as illustrated by the rainfall exceedance probabilities in Figures 6c–e and 7c–e.

Figure 17 complements the objective verification of the sensitivity analysis (Section 4.3) by comparing the reanalysis-driven probabilities of flash flood occurrence for different training approaches against the baseline computed with the full training dataset (Figure 16a). TA1-1 (Figure 17a) and TA3-2 (Figure 17g) produce probability distributions across both western and eastern CONUS that remain closest to the baseline. This is consistent with the objective verification in section 3.4, with ROC curves (Figures 15a and 15g) and PR curves (Figures 15h and 15n) similar to baseline (Figures 12d and 12h), and reliable forecasts below 20% (Figures 15o and 15u), although the reliability diagram shows that probabilities greater than 10% are less frequent than in the baseline (Figure 12l). TA2-2 (Figure 17e) yields a similar spatial structure of areas at risk over the eastern CONUS (blue circle), although with reduced probabilities. The systematic absence of flash flood reports over the western CONUS during training causes the model to suppress probabilities to near-zero over that region (red circle). This spatial asymmetry is consistent with TA2-2's moderate ROC curve (Figure 15e) and reliable forecasts only up to approximately 20% (Figure 15s): the model retains skill where it was trained with positive labels but cannot generalise to the unlabelled western domain. TA1-2 (Figure 17b) still detects the Storm Ida signal but with substantially reduced probabilities and spatial extent, matching its compressed ROC curve (AROC = 0.588; Figure 15b) and narrowed reliability range (Figure 15p). TA1-3 (Figure 17c), TA2-1 (Figure 17d), and TA3-1 (Figure 17f) show severely reduced probabilities across the domain. Among these, TA2-1 (Figure 17d) exhibits the poorest performance, with probability values diminishing to near-zero throughout the entire domain — consistent with its near-diagonal ROC curve (AROC = 0.500; Figure 15d) and almost non-existent reliability diagram (Figure 15r).

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

5.2 Assessment of global forecasts outside the CONUS domain: Valencia flash floods

Valencia (Spain) and adjacent inland regions (Albacete, Cuenca, and Málaga) experienced a prolonged period of intense rainfall between 28 October and 4 November 2024. On 29 October, a cut-off low-pressure system, referred to locally as a "Dana", generated stationary mesoscale circulations that advected moist Mediterranean air onto the eastern coast of Spain over several hours [7]. Rainfall totals broke national records at multiple accumulation periods: 184.6 mm in 1 hour, 620.6 mm in 6 hours, and 720.4 mm in 12 hours. At Turís Mas de Calabarra, 771.8 mm were recorded in 24 hours (Figure 18a), a value second only to the 817.0 mm observed at Oliva (Valencia) in 1987 (Gascón et al., 2025). The torrential rainfall produced widespread surface runoff and riverine flash flooding across the province of Valencia. The most severe impacts concentrated along the small (∼380 km^2) Rambla del Poyo catchment, whose downstream city of Valencia was among the hardest-hit areas despite receiving comparatively little direct precipitation. The event caused 232 fatalities, displaced thousands, and inflicted an estimated 16.5 billion in damage.

The ERA5-ecPoint reanalysis for 29th October 2024 shows that probabilities of exceeding the 1-year threshold reached 70% west of Valencia, particularly over the upstream area of the Rambla del Poyo catchment, and up to 50% over the city itself (Figure 18b). The 50-year return period exceedance probabilities (Figure 18c) provide greater spatial precision, isolating the Valencia region, Cuenca, Albacete, and Málaga as the areas that ultimately experienced the most severe rainfall. Notably, the 50-year exceedance probabilities over the city of Valencia are lower than those upstream (5% instead of 20%), reflecting the orographic enhancement of rainfall over the inland catchments. The data-driven flash flood model's reanalysis-driven predictions (Figure 18d) broadly correspond to the areas identified by the rainfall exceedance maps but provide an important additional signal: the model propagates flash flood risk downstream from the Rambla del Poyo catchment to the city of Valencia, providing a 5 to 10% probability of flash flood occurrence over Valencia (Figure 18c). By incorporating rainfall probabilities in adjacent grid-boxes alongside those in the target grid-box (Section 3.3), the model captures the hydrological connectivity between upstream precipitation and downstream flooding. This downstream propagation of flash flood risk represents one of the model's added values over raw rainfall exceedance probabilities, as it is able to represent more accurately the hazard that ultimately materialised in the city. The forecast evolution of flash flood occurrence probabilities shows a consistent signal with elevated probabilities of flash flood occurrence over Valencia from day 1 (Figure 18e) to day 5 (Figure 18g). The rainfall exceedance signal remained consistent up to day 5, although with decreasing probabilities at longer lead times (not shown). This sustained predictability at medium range is consistent with the synoptic-scale forcing of the DANA system, whose large-scale dynamics are better resolved by the global NWP model beyond the typical predictability limits of convective-scale events.

[7] For more detailed information about the event, please refer to the following AEMET report (in Spanish): www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/estudios/informe_episodio_dana_29_oct_2024_.pdf

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

6 Discussions

6.1 Model architecture under imbalanced datasets: balancing detection sensitivity against false alarm rates

XGBoost gradient boosting achieved the overall best performance for medium-range flash flood prediction from severely imbalanced datasets (full comparative results across all six model implementations are provided in the Supplemental Material). The feed-forward neural network demonstrated no systematic superiority over tree-based methods, whilst requiring approximately 40 times longer training time per fold (Supplemental Material). The hyperparameter importance analysis revealed that, for the classification problem analysed here, simple architectures are preferred: maximum tree depth and learning rate consistently dominate for gradient boosting, whilst neural networks perform well with a single hidden layer. Complex, deep architectures are therefore unnecessary to predict flash flood occurrence. This finding offers considerable operational advantages. Computationally inexpensive models can be retrained more frequently to incorporate new observations, run in ensemble mode to quantify prediction uncertainty, and be deployed at higher spatial resolutions without prohibitive computational cost.

Weighted loss functions (Supplemental Material) increase hit rates from approximately 40% to 90% for gradient boosting. However, this enhanced detection incurs false alarm rates exceeding 50%, a twenty-fold increase compared to balanced configurations. For flash flood early warning systems, this trade-off has direct operational consequences: excessive false alarms erode public trust and reduce response compliance (Emerton et al., 2016), undermining the very purpose of the warning system. Frequency bias and reliability diagrams confirm that balanced configurations yield more reliable predictions (Supplemental Material). For flash flood occurrence prediction, where public trust and sustained response compliance are critical (Bazo et al., 2019), the conservative approach of balanced loss functions is preferable despite lower identification rates.

The minimal degradation between training (2001–2020) and verification (2021–2024) datasets demonstrates good model generalisation, indicating that the learned patterns capture fundamental hydro-meteorological relationships rather than dataset-specific anomalies. Forecast skill degrades with lead time, particularly after day 3 in the reliability diagrams. However, the probabilities at which forecasts remain reliable correspond to the most frequently occurring events, preserving operational utility across the medium range.

6.2 Physical interpretability: reconciling SHAP-derived feature importance with hydrological understanding

The SHAP analysis corroborates the well-documented role of rainfall as the dominant predictor of flash flood occurrence (Schumacher, 2017). Secondary features reveal both expected modulating effects and unexpected correlations that warrant careful interpretation.

Consistent with hydrological understanding, saturated soils (PMSS) amplify flash flood probability for a given rainfall amount (Grillakis et al., 2016). Surprisingly, it appears to be the fourth variable most important in defining flash flood occurrence, behind vegetation coverage and orography steepness. In several flash flood susceptibility analysis, soil moisture is instead one of the primary modulation features (Zhao et al., 2025b; Singh et al., 2021; Zhai et al., 2018).

Vegetation coverage, represented by the leaf area index (LAI), emerges as the second most influential predictor. This ranking is unusual: in flash flood susceptibility studies, vegetation parameters typically rank lower than topography and soil moisture

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

(Costache et al., 2020). Moreover, susceptibility studies generally associate reduced vegetation coverage (low LAI) with increased flash flood risk, as bare or sparsely vegetated surfaces promote rapid surface runoff. In this study, the relationship is reversed: flash flood probability increases with high LAI values. This counterintuitive result likely reflects a seasonal *correlation* that should not be mistaken for *causation*. Both LAI and flash flood occurrence peak during summer months across the CONUS (Dougherty and Rasmussen, 2019): vegetation reaches maximum density precisely when convective rainfall, and hence flash flood frequency, is highest. The model thus appears to exploit LAI as a temporally informative proxy for the flash flood season rather than as a physically meaningful predictor of catchment susceptibility. If confirmed, this confounding effect would need to be addressed, for example by replacing LAI with a de-seasonalised vegetation anomaly or by incorporating an explicit seasonality feature, to isolate real vegetation effects on flash flood occurrence from mere seasonal co-occurrence.

As the LAI, the surprise around the physical interpretation regarding the topographic slope (SDFOR) is twofold. SDFOR ranks third in importance, whilst many susceptibility studies show that orography steepness is one the main flash-flood-generating features (Luo et al., 2025; Zhao et al., 2025c; Chen et al., 2019). Moreover, topographic steepness, shows an inverse relationship to physical expectations: SHAP values indicate that flatter terrain increases flash flood probability. As with LAI, this counterintuitive association is better understood as a *correlation*, driven by the spatial distribution of impact reports, rather than a genuine *causal* relationship. Whilst runoff generation concentrates in steep headwater catchments, flash flood impacts and the reports that document them concentrate in downstream valleys and urban areas where populations reside (Marjerison et al., 2016). The Storm Event Database thus over-represents flash flood occurrence in flat, densely populated terrain (Section 2.2). A model trained on discharge observations rather than impact reports would likely yield different, physically more consistent feature importance rankings for SDFOR.

These findings underscore that SHAP values reflect statistical associations learned from the training data, including its biases, and should not be interpreted as evidence of physical causation without independent corroboration.

6.3 Extending regionally trained models to data-sparse domains: the primacy of hydro-climatic diversity over global coverage

Training exclusively on a data-rich region with sufficient hydro-climatological diversity (TA3-2) enables successful extrapolation to ungauged areas. Despite the SHAP outcomes shown in the previous section, this study establishes that XGBoost was able to learn generalisable hydro-meteorological patterns from representative training data and preserves probability distributions when applied to unseen domains. This result aligns with Kratzert et al. (2024), who demonstrated similar transferability for riverine flood prediction using long short-term memory networks trained on the CAMELS dataset. The implication for operational practice is significant: rather than developing numerous local models that struggle beyond their training domain (Santos et al., 2025), a single model trained on a sufficiently diverse and densely observed region can be deployed globally, provided the training domain spans the range of hydro-climatic conditions encountered during inference. The Valencia case study empirically validates this strategy, demonstrating that TA3-2 successfully transfers learned patterns to a region with distinct hydro-climatic characteristics and produces skilful forecasts up to day 5. Where better local observations become available, such a foundational model could be fine-tuned regionally to further improve performance.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Alternative training approaches using sparse global data (TA1) or regionally limited observations that leave entire domains unlabelled (TA2) degrade performance substantially. TA1 demonstrates that full-domain but sparse coverage produces progressive deterioration in both reliability and discrimination as observational density decreases. At 90% data reduction (TA1-3, which is the closest representation of what it would be training a global model with today's global impact databases such as EMDAT and Desinventar), model performance collapses to forecasts with an unusable performance. TA2-1 yields similar failure, as limited hydro-climatic diversity and lower flash flood frequency in the western CONUS provide insufficient training signal. Reliability diagrams show that models trained under TA1-3 and TA2-1 struggle to produce probabilities different from zero, exhibiting behaviour consistent with the model effectively abandoning attempts to predict the minority class. The precision-recall curves reveal a further diagnostic pattern. Models trained with fewer positive reports terminate at low recall values but often achieve higher precision at those low recall values than models trained on more comprehensive datasets. This arises because extreme data scarcity forces the model to become hyper-conservative, triggering positive predictions only under conditions that almost exactly replicate its limited training examples. The resulting selectivity yields high precision for a negligible number of predictions whilst missing most positive cases. In contrast, models trained with more diverse positive examples learn richer representations that generalise across varied flash flood contexts; though precision at very low recall values is lower, these models maintain reasonable precision across operationally meaningful recall levels. For life-safety applications, detecting hazards across their full diversity outweighs perfect precision on a negligible fraction of events. These results carry direct implications for the use of global impact databases such as EM-DAT (Panwar and Sen, 2020) for training data-driven flash flood models: their sparse and spatially biased coverage would likely produce analogous performance degradation.

6.4 From research to operations: remaining challenges

The transition from this research demonstration at ERA5's 31 km resolution to operational deployment at a resolution more amenable to flash flood prediction (e.g., 10 km) presents computational and logistical challenges. Whilst inference is computationally inexpensive, retraining at finer resolution would demand substantially greater computational resources. Model updating introduces further complexity: continuous integration of new observations requires automated quality control pipelines and regular retraining cycles that demand both computational infrastructure and human oversight.

The absence of comprehensive global impact observations prevents rigorous verification at local scales outside well-monitored regions. Global databases such as EM-DAT and DesInventar exhibit severe reporting biases (Panwar and Sen, 2020), capturing only events exceeding mortality or economic thresholds whilst under-representing smaller but locally significant flash floods. Spatial biases compound this limitation, with reporting quality correlating strongly with institutional capacity, media coverage, and proximity to urban centres, precisely inverse to the distribution of vulnerability (Panwar and Sen, 2020; Gaume et al., 2009). These constraints necessitate reliance on case studies for the verification of global forecasts, which, whilst informative, do not establish general performance statistics over the full global domain. Future solutions may emerge through non-traditional observation sources: satellite-detected flooding as a proxy for ground impacts, crowdsourced and social media observations (though extracting reliable information requires advanced natural language processing and verification protocols),

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

and IoT devices such as low-cost weather stations and water level sensors that could expand coverage in currently unmonitored regions.

6.5 Beyond flash floods: broader implications and future directions

The transfer learning framework demonstrated here addresses a fundamental challenge shared across hydro-meteorological hazards: dense, quality-controlled observations exist only in specific regions, yet predictions are needed globally. Lightning detection, for instance, benefits from ground-based networks in North America and Europe, yet vast regions across Africa, Asia, and South America remain unmonitored despite experiencing intense thunderstorm activity. Landslide inventories exhibit an analogous pattern, with detailed historical records in specific countries whilst global databases remain sparse and inconsistent. The framework established here could be adapted to these hazards: models trained on well-observed regions could transfer learned relationships and provide predictions in regions currently lacking warning systems due to insufficient observational databases.

The demonstrated transferability across diverse hydro-climatic regions carries implications for climate change adaptation, though model robustness under non-stationary conditions requires further investigation. As global warming intensifies the hydrological cycle (IPCC, 2023), regions historically experiencing infrequent flash flooding may transition into high-risk zones, whilst traditional flood-prone areas may face unprecedented extremes (Fowler et al., 2021). Models trained on climatologically diverse regions could provide anticipatory warnings for communities entering unfamiliar climate regimes. Critical questions remain, however: how frequently must models be retrained to capture evolving precipitation-runoff relationships, and can they predict future extremes exceeding historical analogues? Preliminary evidence from Bertola et al. (2023) suggests positive signals for model robustness under changing conditions, though flash-flood-specific verification under non-stationary climates is still needed. Running the model on retrospective reanalysis data could generate historical flash flood climatologies, enabling assessment of changing frequency patterns and improved understanding of triggering hydro-meteorological conditions.

7 Conclusions

This study provides the first empirical evidence that global, medium-range forecasts of flash flood occurrence are achievable using parsimonious data-driven models and readily available data. A simple XGBoost gradient boosting model, trained on regional flash flood impact reports and driven by global reanalysis and NWP forecasts, produces skilful and reliable probabilistic predictions up to day 5 over the CONUS. A spatially constrained sensitivity analysis further demonstrates that a model trained exclusively on a hydro-climatologically diverse and observation-dense sub-domain generalises to unseen regions, including outside the training continent, as validated by the Valencia flash floods case study. This transferability result has direct implications for operational practice: rather than requiring dense observational networks everywhere, global flash flood forecasts can be underpinned by high-quality regional observations, provided the training domain spans the range of hydro-climatic conditions encountered during deployment. Collectively, these results close one of the most persistent gaps in operational hydrology,

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

the absence of flash flood forecasts over a continuous global domain at medium-range lead times, and offer a concrete pathway towards the UN's "Early Warnings for All" target (UN, 2022).

Several directions for future work follow from the limitations identified in this study. First, the current model operates at ERA5's 31 km resolution, which limits its capacity to resolve the small catchments (typically below 500 km^2) where flash flood impacts concentrate; transitioning to a finer grid, for example 10 km, is a prerequisite for operational deployment and would require both higher-resolution input data and substantially greater computational resources for training. Second, model skill for storm-scale convective systems, which exhibit lower predictability than synoptically forced events at medium-range lead times, remains limited; incorporating convection-permitting NWP outputs or ensemble-based representations of convective uncertainty could improve predictions for this class of events. Third, the SHAP analysis revealed that the model's learned associations between flash flood occurrence and both vegetation coverage (LAI) and topographic slope (SDFOR) reflect correlations in the training data rather than causal physical relationships; future iterations should investigate deseasonalised vegetation indices, alternative topographic descriptors, or the inclusion of additional predictors (e.g., urban imperviousness, land use classification) to disentangle genuine physical drivers from artefacts of the observational database. Fourth, the model is currently trained on a single observational source; integrating complementary observation types, such as satellite-detected flooding, crowdsourced reports, and discharge records from small catchments where available, could both enrich the training signal and enable more robust verification over the global domain. Fifth, the question of model robustness under non-stationary climatic conditions warrants systematic investigation, given that intensification of the hydrological cycle (IPCC, 2023) may shift both the frequency and the spatial distribution of flash flood occurrence beyond historical analogues; retrospective application of the model to reanalysis data spanning several decades could provide a first assessment of whether the learned relationships remain stable under evolving hydro-climatic regimes.

Appendix A: Post-processing of the flash flood impact reports

To assess the performance of the data-driven flash flood occurrence forecasts, these must be evaluated against ground-truth observations, which in this study are provided by flash flood impact reports in the NOAA's Storm Event Database (refer to Section 2.2 for the detailed description of the database).

Alongside events classified as "Flash Flood" (EVENT_TYPE in Table A1), events classified as "Heavy Rain", "Hurricane/Typhoon", and "Tropical Storm" whose FLOOD_CAUSE was "Heavy Rain" or "Heavy Rain/Snow Melt" (Table A1) were also retained because they frequently produce flash floods but are catalogued under separate categories in the Storm Events Database. This brought the initial count of considered flash flood reports from 86,480 to 119,797 (39% increase), from 1950 to 2024. Any reports lacking valid geographical coordinates (BEGIN_LAT, END_LAT, BEGIN_LON, and END_LON in Table A1) or reporting timestamps (BEGING_DATE_TIME and END_DATE_TIME) were discarded, so only 96,989 raw point reports were retained, from 2001 to 2024. To account for multi-day events (where BEGING_DATE_TIME $\neq$ END_DATE_TIME), records spanning more than one calendar day were expanded such that each day within the event's reported duration generated an individual entry. The resulting dataset comprised one row per event-day, retaining all the meta-

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table A1. Summary of the key fields extracted from the NOAA Storm Events Database for flash flood events.

Key*	Type	Units	Description	Origin
EVENT_ID	String	–	ID assigned by NWS for each individual storm event contained within a storm episode.	Original
STATE	String	–	The (spelt out) name of the state where the event occurred.	Original
CZ_TIMEZONE	String	–	Time zone for the County/Parish, Zone or Marine Name (e.g., EST, CST, MST).	Original
SOURCE	String	–	The source reporting the weather event (e.g., Public, Newspaper, Law Enforcement, Broadcast Media, ASOS, Trained Spotter, CoCoRaHS).	Original
EVENT_TYPE	String	–	Type of event (e.g., riverine or flash flood). Permitted types are listed in Table 1 of Section 2.1.1 of the NWS Directive 10-1605.	Original
FLOOD_CAUSE	String	–	Reported or estimated cause of the flood (e.g. Ice Jam, Heavy Rain, Heavy Rain/Snow Melt).	Original
BEGIN_DATE_TIME	Date and Time	UTC	Date and time of the beginning of the flash flood event (MM-DD-YYYY hh:mm:ss, 24-hour format).	Original
END_DATE_TIME	Date and Time	UTC	Date and time of the end of the flash flood event (MM-DD-YYYY hh:mm:ss, 24-hour format).	Original
BEGIN_LAT	Float	Decimal degrees	Latitude of the begin point for the event or damage path.	Original
BEGIN_LON	Float	Decimal degrees	Longitude of the begin point for the event or damage path.	Original
END_LAT	Float	Decimal degrees	Latitude of the end point for the event or damage path.	Original
END_LON	Float	Decimal degrees	Longitude of the end point for the event or damage path.	Original
REPORT_DATE	Date	UTC	Calendar date assigned to each day spanned by a multi-day event, generated by expanding records between BEGIN_DATE_TIME and END_DATE_TIME.	Derived

* For the description of the database fields not considered in this study, the reader is referred to:
https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Bulk-csv-Format.pdf.

data from the original database but with a different date, stored as a new field (REPORT_DATE, Table A1). The final count of expanded point flash flood reports considered in the study is 108,555: 86,844 point flash flood reports were used to train the ML model (training period: 2001 to 2020), whilst 21,711 point flash flood reports were used for independent model validation (verification period: 2021 to 2024).

The data-driven flash flood occurrence forecasts and the flash flood impact reports are not directly comparable: the forecasts are defined on the ERA5's grid and are accumulated over a 24-hourly period, whilst the raw impact reports are provided at points (where BEGIN_LAT = END_LAT and BEGIN_LON = END_LON) or polygons (where BEGIN_LAT $\neq$ END_LAT and BEGIN_LON $\neq$ END_LON) for a specific instantaneous timestamp (REPORT_DATE), as described above and shown

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

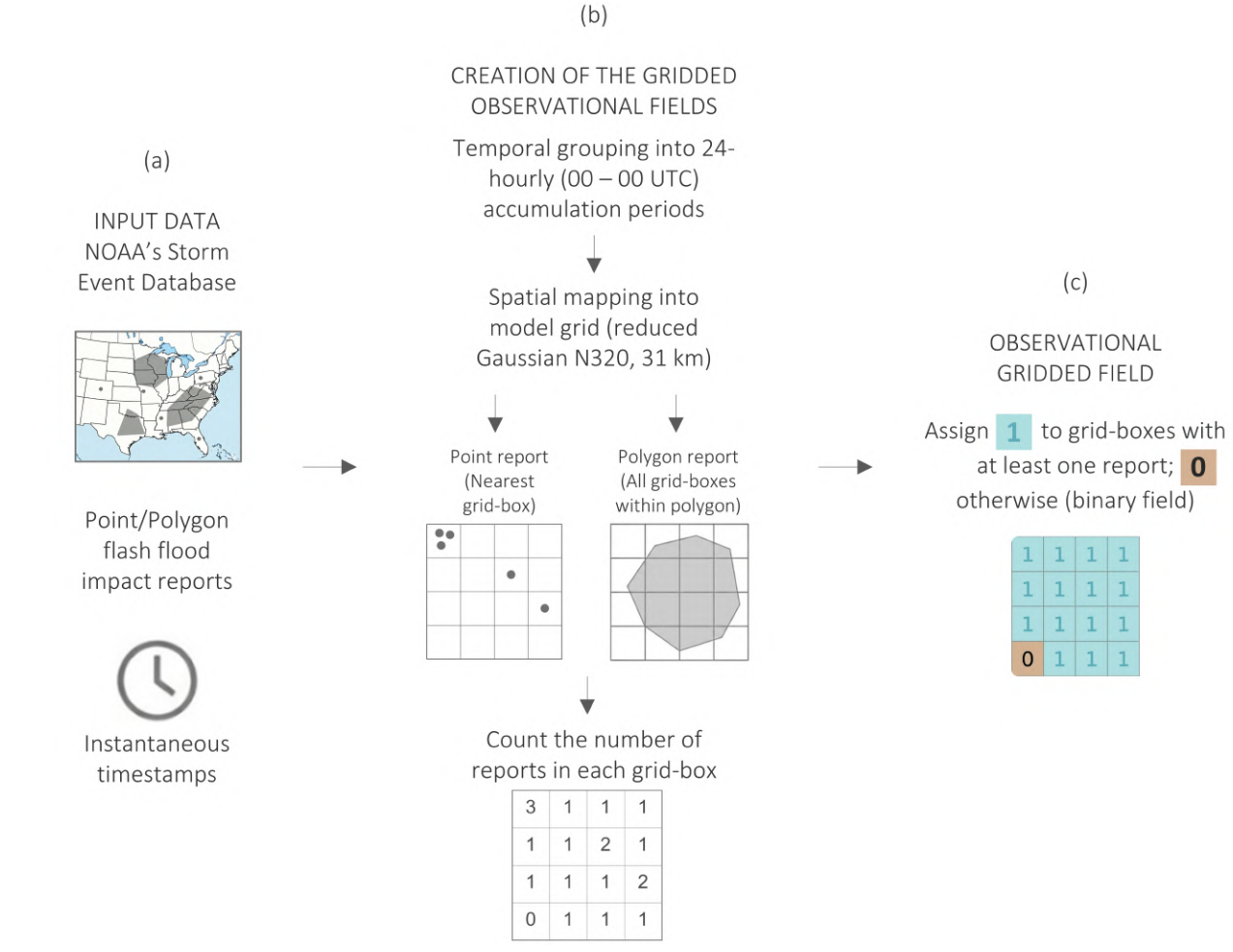

Figure A1. Schematic about how point/polygon flash flood impact reports with instantaneous timestamps were post-processed into gridded, accumulated observational fields. Panel (a) shows the input flash flood impact reports from NOAA's Storm Event Database, consisting of point/polygon flash flood impact reports with instantaneous timestamps. Panel (b) shows the logic followed for the post-processing: (1) grouping the instantaneous reports into 24-hourly accumulation periods ending at 00 UTC, and (2) mapping them into the considered grid by assigning point reports to the nearest grid-box and polygon reports to all the grid-boxes within the polygon. Panel (c) shows how the total number of impact reports in each grid-box is then transformed into a binary field by assigning 1s to grid-boxes containing at least one report and 0s otherwise.

in Figure A1a. Hence, the impact reports must be converted to the same grid and accumulation period as the forecasts (Figure A1b). Practically, impact reports are first grouped into corresponding 24-hour accumulation periods (00–00 UTC). Each point/polygon report is then mapped to the model grid: point reports are assigned to the nearest grid-box, whereas polygon reports are assigned to all the grid-boxes within the polygon. The total number of reports accumulated in each grid-box is then counted per accumulation period (Figure A1b). All the grid-boxes containing at least one flash flood report are assigned the

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Figure A2. Comparison of point and gridded flash flood impact reports. (a) Daily timeseries point (red) and gridded (black) flash flood report counts, accumulated over 24-hourly periods ending at 00 UTC, for 2021. The blue circle highlights the report counts for the 24-hourly accumulation period ending on 2021-09-02 at 00 UTC (Storm Ida). (d) Spatial distribution of point and gridded reports for the same day; the inset focuses on the New York area.

value 1; 0, otherwise (Figure A1c). Spatio-temporal buffers were not applied to the impact reports, unlike other studies using forecasts with finer spatial and temporal resolutions (Cavaiola et al., 2024). This choice is unlikely to affect the results as only 0.1% of all reports lie sufficiently close to a grid-box boundary or the 24-hour accumulation cut-off that omitting a buffer could plausibly shift an event into a neighbouring grid box or an adjacent accumulation period. Hence, no additional adjustments were made to account for uncertainties in the reports' location or timing.

Figure A2a illustrates the daily time series of point and gridded report counts for 2021, showing that gridded counts are consistently lower because multiple point reports within a single grid-box are collapsed to a binary value of 1. This binarisation prioritises flash flood occurrence over frequency, ensuring that differences in predicted probability reflect the predictability of the triggering hydro-meteorological system rather than social factors involved when people report extreme hydro-meteorological events that may reflect the nature of the triggering event (e.g., large-scale events such as hurricanes or extra-tropical storms

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table B1. ERA5 parameters used as model features in this study.

Name Parameter	Symbol	Range of values	Units	Mars ID	Type	Accumulation
Volumetric soil water, layer 1 (0–7 cm)	swvl1	Float, >0	$m^3\ m^{-3}$	39	Dynamic	Instantaneous
Volumetric soil water, layer 2 (7–28 cm)	swvl2	Float, >0	$m^3\ m^{-3}$	40	Dynamic	Instantaneous
Volumetric soil water, layer 3 (28–100 cm)	swvl3	Float, >0	$m^3\ m^{-3}$	41	Dynamic	Instantaneous
Soil type	slt	Integer, 0 to 7*	–	43	Static	n/a
Std. dev. of filtered sub-grid orography	SDFOR	Float, >0	m	74	Static	n/a
Slope of sub-grid orography	slor	Float, 0 to 1 (0–90°)	–	163	Static	n/a
Leaf area index, low vegetation	LAI_lv	Float, 0–7	$m^2\ m^{-2}$	66	Dynamic, climatological	Instantaneous
Leaf area index, high vegetation	LAI_hv	Float, 0–7	$m^2\ m^{-2}$	67	Dynamic, climatological	Instantaneous
Low vegetation cover	cvl	Float, 0–1	–	27	Static	n/a
High vegetation cover	cvh	Float, 0–1	–	28	Static	n/a

* Soil type codes: coarse (1), medium (2), medium fine (3), fine (4), very fine (5), organic (6), and tropical organic (7). For the leaf area index, 0 corresponds to bare soil and 7 to dense canopy. To know how to retrieve these fields, refer to the Mars Catalogue at https://apps.ecmwf.int/mars-catalogue/?class=ea&stream=oper.

typically generate far more reports per event than isolated convective storms) or demographic factors (e.g., events in densely populated areas will naturally produce more reports). Without it, events falling in the latter category (i.e., large-scale events over highly populated areas) would dominate the training signal. The spatial distribution for the peak event of 2021 (i.e., the 24-hour period ending 2021-09-02 at 00 UTC, associated with Storm Ida and highlighted by the blue circle in Figure A2b) illustrates this effect: clustering of point reports is greatest where the event is most widespread and over a highly populated area, yet higher counts do not necessarily indicate a more severe hydrological hazard. Conversely, events in sparsely populated areas may generate fewer reports (e.g., western CONUS in Figure A2b) yet carry comparable hydro-meteorological significance.

Appendix B: Model features

B1 Standard deviation of the filtered sub-grid orography (representing the orographic steepness)

The standard deviation of the orography (SDFOR, Table B) is a static, time-invariant field, representing the statistical variability of terrain elevation at higher spatial resolution (typically at 1 km) than the model grid resolution (for ERA5, ∼31 km). The

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table B2. Maximum soil saturation values for each ERA5 soil type code.

Soil Type Code	**1** Coarse	**2** Medium	**3** Medium Fine	**4** Fine	**5** Very Fine	**6** Organic	**7** Tropical Organic
Maximum Saturation (–)	0.403	0.439	0.430	0.520	0.614	0.766	0.472

parameter is expressed in metres. High values (typically above 100-500 meters) indicate significant topographic heterogeneity within model grid-boxes, representing the presence of valleys, ridges, peaks, and other terrain features that cannot be explicitly resolved at the model grid resolution (ECMWF, 2016).

B2 Percentage of soil maximum saturation (representing antecedent soil moisture)

To estimate the antecedent soil moisture, this study considers the percentage of soil maximum saturation in the top 1 metre layer of soil (ECMWF, 2016), sampled 24 hours prior to the flash-flood-triggering rainfall event. Sampling on the same day would introduce a spurious correlation between soil moisture and rainfall (as the rainfall event itself modifies the soil moisture field), compromising the independence of the two predictors and potentially inflating model skill. The 1 metre depth corresponds to the active "root zone" represented by the first three layers of the ECMWF's IFS land surface scheme, defined as Layer 1, between 0 and 7 cm, Layer 2, between 7 and 28 cm, and Layer 3, between 28 and 100 cm (Table B1). Table B2 shows the pre-defined values of maximum soil saturation used at ECMWF for different types of soil (Balsamo et al., 2009).

Equations B1 and B2 show how the fields containing the percentage of soil saturation were computed:

$$\text{max\_sat\_field} = \sum_{i=01}^{N} \text{max\_sat}_i \, \mathbf{1}_{\{s=c_i\}} \tag{B1}$$

$$\mathbf{1}_{\{s=c_i\}} = \begin{cases} 1, & s = c_i, \\ 0, & s \neq c_i. \end{cases} \tag{B2}$$

where, N goes from 1 to 7, and it represents the soil type codes. Each grid box is assigned a single, dominant soil type. The indicator function $\mathbf{1}_{\{s=c_i\}}$ acts as a binary filter: it returns 1 only when the grid box's soil type s matches a specific category c_i; it returns 0 otherwise.

Equation B3 computes the soil water content (swvl) over the top 1 metre layer of the soil integrating the soil water content over the top three layers:

$$\text{swvl} = \sum_{j=1}^{M} \text{swvl}_j \, \text{depth}_j \tag{B3}$$

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

where M goes from soil layer 1 to 3. The percentage of the soil maximum saturation is considered a dynamic field because the values of swvl in the three soil layers change at every reanalysis and forecast run.

Finally, the equation B4 computes the percentage of soil saturation as follows:

$$\text{swvl\_perc} = \frac{\text{swvl}}{\text{max\_sat\_field}} \tag{B4}$$

B3 Leaf area index (representing vegetation coverage)

The leaf area index corresponds to a non-dimensional number representing the square metres of leaf area per square metre of the earth's surface[8]. It has a value of 0 over bare ground or where there are no leaves, and it grows as the vegetation coverage increases, typically up to values equal to 7. In ERA5, as in the ECMWF IFS, the leaf area index varies only climatologically, month by month. Hence, anomalous weather (e.g. winds stripping leaves from trees or widespread fire damage) has no effect on its value.

The total leaf area index is computed as the coverage-weighted sum of the high- and low-vegetation components (Table B1) as shown by the following equation:

$$\text{LAI}_{\text{total}} = (\text{LAI}_{\text{high}} \times \text{Cover}_{\text{high}}) + (\text{LAI}_{\text{low}} \times \text{Cover}_{\text{low}})$$

Appendix C: Development of ML models under imbalanced training datasets

C1 Ensemble model families

Ensemble learning techniques (Figure 8) manage class imbalance effectively (Ayodele, 2023; Altalhan et al., 2025). Within this category, bagging, boosting, and neural networks were considered due to their effective extraction of complex non-linear patterns from tabular data (Shwartz-Ziv and Armon, 2022). To optimise the model's ability to generalise to unseen data, each family adopts a different strategy to balance the two competing model error sources: bias, measuring the prediction's systematic inaccuracies due to model oversimplification, and variance, measuring the model's sensitivity to fluctuations in the training data (Ranglani, 2024). Bagging primarily reduces variance by training multiple independent models in parallel, each using random training samples. Boosting primarily reduces bias by training weak models sequentially, making them stronger at each step by targeting the errors made by their predecessors. Neural networks consist of interconnected layers of nodes that learn to map inputs to outputs by adjusting connection weights through iterative error minimisation. When properly fine-tuned, neural networks can reduce both bias and variance as deep architectures can capture complex patterns whilst regularisation techniques prevent overfitting. Due to their proven robustness and interpretability in flash flood prediction (Santos et al., 2025), we tested

[8]https://confluence.ecmwf.int/display/FUG/Section+2.1.4.7+Modelling+vegetation.+Leaf+area+index

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table C1. Loss function configurations used to address class imbalance in model training.

Configuration	Abbrev.	Class weighting	Loss formula
Binary cross-entropy	BCE	Equal weight to both classes	$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$
Weighted binary cross-entropy	W-BCE	Optimisable weight w^+ applied to the positive (minority) class (Aurelio et al., 2019; Rezaei-Dastjerdehei et al., 2020).	$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[w^+ y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$

$y_i \in \{0, 1\}$ is the observed label, $\hat{y}_i$ is the predicted probability, N is the number of training samples, and w^+ is the optimisable positive class weight (see Table C2) in the Appendix C3.

the gradient boosting algorithm for boosting, using the popular XGBoost[9] and LightGBM[10], and CatBoost[11] implementations, the random forest algorithm for bagging, using the XGBoost and LightGBM implementations in random forest mode, and feed-forward neural networks, using Keras with TensorFlow backend[12].

C2 Loss functions

Flash floods are rare events, resulting in a substantial class imbalance between positive (yes-events) and negative (non-events) samples in the training set. Two loss function configurations were evaluated to assess whether explicitly compensating for this imbalance improves predictive skill (Table C1). The weighted variant introduces a single additional hyperparameter, w^+, which up-weights the minority class during optimisation. This new hyperparameter is tuned alongside the others (Appendix C3).

C3 Hyperparameter tuning

XGBoost's model performance was optimised via Bayesian hyperparameter tuning over the search space, defined in Table C2. The six hyperparameters govern complementary aspects of the learning process for XGBoost (i.e., model capacity, regularisation, and class imbalance handling), and were tuned jointly to minimise overfitting while preserving sensitivity to the minority flood class. Under the standard BCE configuration, `scale_pos_weight` was held fixed at 1.0, reducing the effective search space to five dimensions.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Table C2. Hyperparameter search space for the XGBoost gradient boosting implementation.

Hyperparameter	Type	Search range	Role in model training
`n_estimators`	Integer	[100, 500]	Number of sequential boosting rounds (trees); higher values increase model capacity but risk overfitting and lengthen training time.
`max_depth`	Integer	[3, 10]	Maximum depth of each decision tree; controls the complexity of individual learners and the bias–variance trade-off.
`learning_rate`	Float	[0.01, 0.3]	Shrinkage factor applied to each tree's contribution; lower values require more estimators but improve generalisation.
`subsample`	Float	[0.6, 1.0]	Fraction of training samples drawn (without replacement) to fit each tree; reduces variance by introducing stochasticity.
`colsample_bytree`	Float	[0.6, 1.0]	Fraction of features randomly selected when constructing each tree; reduces inter-tree correlation and variance.
`scale_pos_weight*`	Float	[1.0, 10.0]	Multiplicative weight assigned to positive-class samples; compensates for class imbalance by increasing the loss contribution of minority-class errors.

*The `scale_pos_weight` parameter is tuned only under the weighted binary cross-entropy (W-BCE) loss configuration and is fixed at 1.0 under standard BCE. All other parameters are tuned under both loss configurations.

C4 Objective verification

C4.1 Defining forecast yes-events

A grid-box is classified as a yes-event if the probability of flash flood occurrence exceeds a certain probability threshold. In this case, the grid-box is assigned the value 1 (yes-event); 0 otherwise (non-event), resulting in a forecast binary field (Figure C1). Conventionally, a probability threshold of 50% is generally applied. However, for rare events such as flash floods, this would yield very few yes-event classifications, rendering verification insensitive to model performance. The threshold is therefore optimised on the F1-score, which balances precision and recall by penalising false alarms and missed events equally (Hancock et al., 2022). This approach identifies the decision boundary that maximises predictive skill given the inherent trade-off between detection rate and false alarm rate.

C4.2 Defining the probabilistic contingency table

To construct the probabilistic contingency table, one must derive the distribution of observed yes- (blue distribution in Figure C2a) and non-events (distribution in pink). Second, a probability threshold for exceeding the verifying rainfall threshold (dotted-dashed vertical black line in Figure C2a) partitions the rainfall forecasts into predicted yes-events (to the right of the

[9]https://xgboost.readthedocs.io/en/stable/tutorials/rf.html
[10]https://lightgbm.readthedocs.io/en/latest/index.html
[11]https://catboost.ai/docs/en/
[12]https://www.tensorflow.org/guide/keras

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

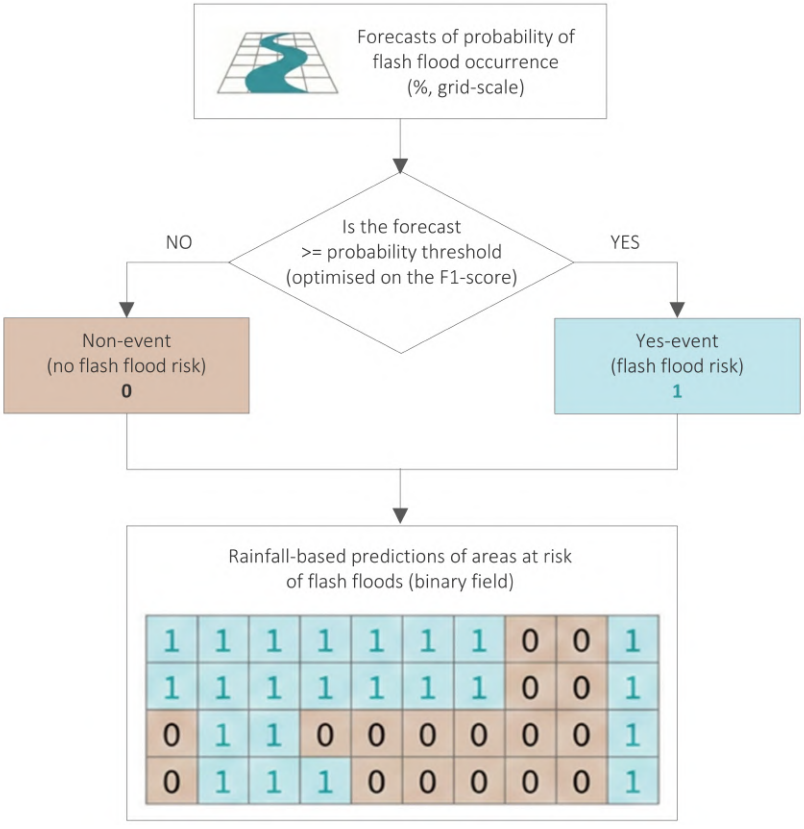

Figure C1. Schematic on how yes- and non-events are defined for probability (%) of flash flood occurrence forecasts. The forecasts in each grid-box are compared against a probability threshold (optimised on the F1-score). If the forecast exceeds or is equal to the threshold, the grid-box is classified as a "yes-event" (value 1, shown in light green), indicating a risk of flash flood. Conversely, forecasts below the threshold are classified as "non-events" (value 0, shown in light brown). The resulting output is a binary field representing the predictions of areas at risk.

threshold) and predicted non-events (to the left). This partitioning yields four categories: hits (H, green diagonal hatching) — events correctly predicted and observed; false alarms (FA, pink shading) — events predicted but not observed; misses (M, brown shading) — events observed but not predicted; and correct negatives (CN, orange horizontal hatching) — non-events correctly predicted. Third, these counts populate a 2×2 contingency table for that threshold (Figure C2b). In practice, the 2×2 contingency table is populated by comparing co-located grid boxes in the predicted (Figure C1) and observed fields (Figure A1). A hit occurs when corresponding grid-boxes are assigned a value of 1; a correct negative when both are 0; a miss occurs when the observed value is 1, but the forecast is 0, and a false alarm occurs when the forecast is 1, but the observed value is 0 (Figure C2c). By repeating this process across all considered probability thresholds, a series of 2x2 contingency tables is obtained, one per threshold.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

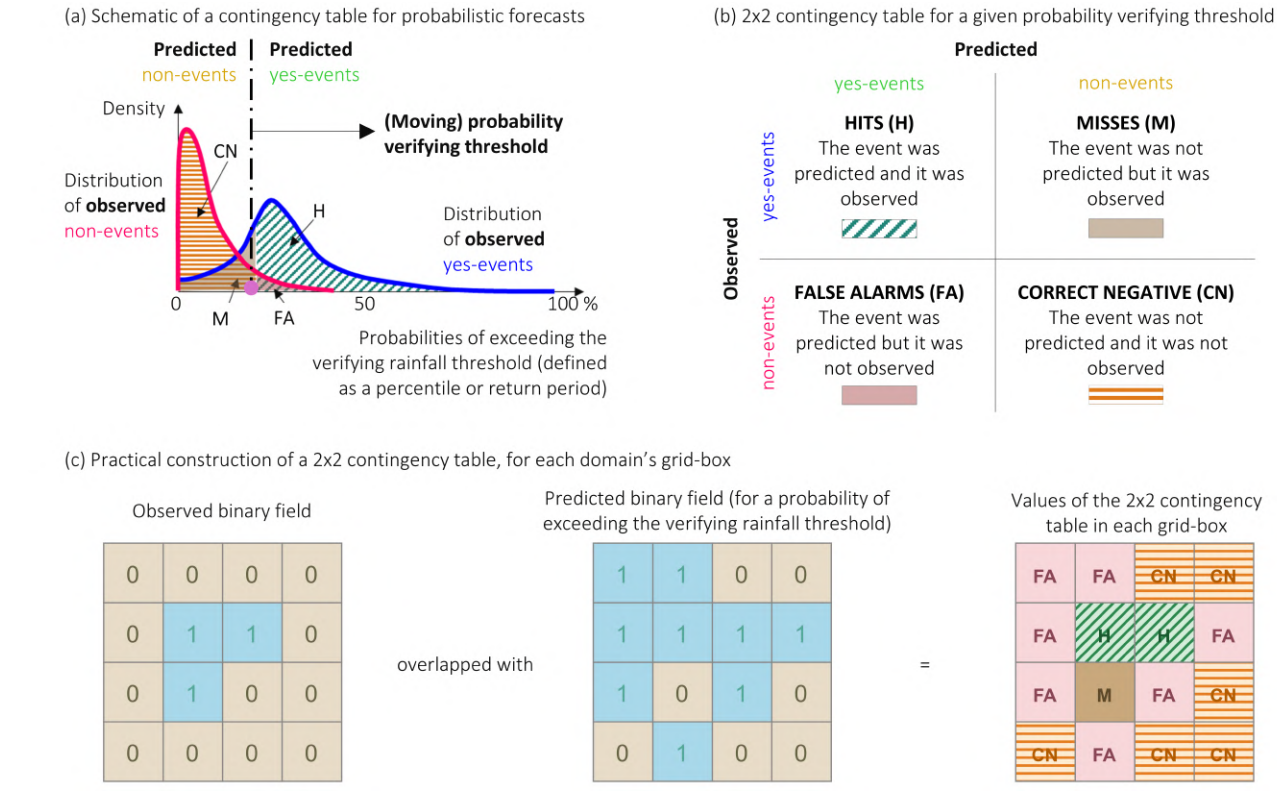

Figure C2. Contingency table for probabilistic forecasts. (a) Schematic on how a contingency table is built for probabilistic forecasts with a moving probability verifying threshold. (b) 2x2 contingency table when fixing the probability verifying threshold. (c) Schematic of the practical construction of the 2x2 contingency table for each grid-box in the considered geographical domain.

Unlike stationary observations, such as instruments installed at a specific location (e.g., rain gauges or discharge gauges - that provide a continuous timeseries of observed yes- *and* non-events), impact reports are non-stationary observations that record only observed yes-events. In the first case, all four quadrants of the contingency table can be quantified. Since in the latter case (when using impact reports as ground truth) it is impossible to answer the question "if there are no reports at a location, is it because an event happened but nobody reported it, or because there was no event to report?", it is more difficult to fill all four quadrants of the contingency table. Some studies using impact reports as ground truth verify only yes-events, with the caveat that only quadrant H (i.e. hits) and M (i.e. misses) of the contingency table can be populated (Robbins and Titley, 2018). This approach offers only a partial assessment of the forecasts' performance as it does not attempt to quantify FAs. In this thesis, the approach developed by Tsonevsky et al. (2018) and Pillosu et al. (2024) is adopted instead. They assume that a non-report represents an observed non-event. This assumption is acceptable for the considered impact reports, as they undergo an acceptable quality control. Nonetheless, given the constraints of the observational data, this approach will

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

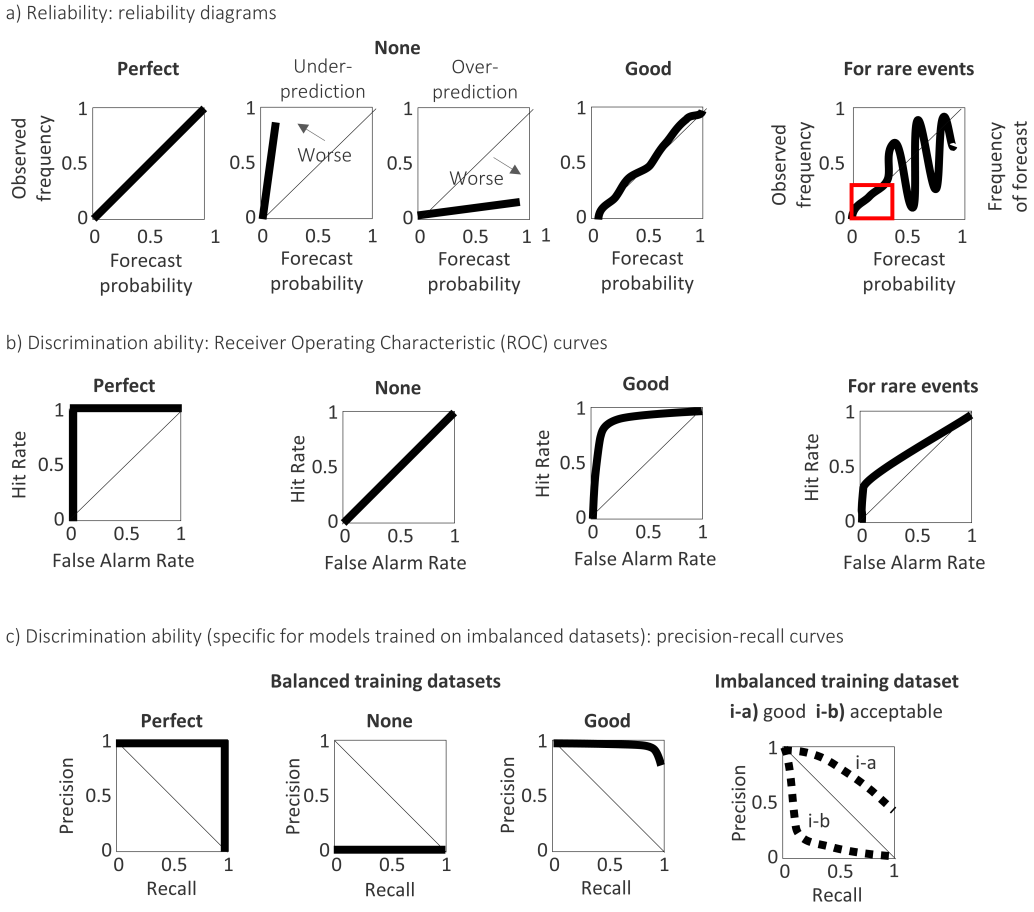

Figure C3. Breakdown verification scores. (a) Examples of reliability diagrams for forecasts with perfect, none, and good reliability, and for rare events. The red square indicates the forecast probabilities with the largest number of cases because, outside the red square, the reliability diagrams become noisy. (b) Similar to (a), but for PR curves. (c) Examples of PR curves for perfect, none, and good discrimination ability. The PR curves for forecasts trained on imbalanced datasets are also shown, with good (i-a) and acceptable (i-b) discrimination ability.

775 inherently inflate the number of FAs. However, this approach will provide a broader and more trustworthy evaluation of the rainfall-based predictions of areas at risk of flash floods.

C4.3 Assessing reliability

The frequency bias (FB) assesses the overall reliability of the rainfall-based predictions of areas at risk of flash floods. The frequency bias represents the fraction of the total number of predicted yes-events over the total number of yes-events in the
780 observations. It is calculated with the equation C1:

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

$$\text{Frequency Bias} = \frac{\text{H} + \text{FA}}{\text{H} + \text{M}} \tag{C1}$$

FB values range from 0 to $+\infty$, with FB = 1 indicating perfect bias. Values greater or smaller than 1 indicate, respectively, over- and under-prediction of the observed yes-events. It is worth noting that FB measure the overall ratio of forecast events to observed events and is not a measure of forecast skill. As such, it can provide a score of 1 when there are compensating errors. Moreover, the FB might show large overestimations if the observed event is heavily underreported, as it is in our case.

Reliability diagrams are used instead as breakdown scores to assess reliability (Figure C3a). They plot the relative forecast probabilities of an event against its corresponding relative observed frequency, indicating how reliable the forecast probabilities are at different probability classes. For perfect forecasts, when the forecasts show x% probability of occurrence, observations should meet the criteria x% of the time, so that the reliability curve lies on the diagonal. If the reliability diagram is above the diagonal for a specific forecast probability, those forecasts are under-predicting the likelihood of observing a yes-event. If it lies below the diagonal, there is over-prediction. When analysing reliability diagrams, especially when considering rare events, it is important to know the frequency distribution of forecasts issued for specific probabilities. For example, the small probability thresholds are the most important (within the red square in C3a). The sample of forecasts with high probabilities (outside the red square) will be rather small, and the reliability diagram is likely to appear noisy. Dimitriadis et al. (2021) propose a formulation for more stable reliability diagrams in case of rare events, but this formulation will not be considered in this study.

C4.4 Assessing discrimination ability

The Relative Operating Characteristic (ROC) curve is built from the probabilistic contingency table in Figure C3b, mapping Hit Rates (HR) against False Alarm Rates (FAR), computed from equations C2 and C3, respectively:

$$\text{HR} = \frac{\text{H}}{\text{H} + \text{M}} \quad [\text{values between 0 and 1}] \tag{C2}$$

$$\text{FAR} = \frac{\text{FA}}{\text{FA} + \text{CN}} \quad [\text{values between 0 and 1}] \tag{C3}$$

HRs are mapped (Y-axis) against FARs (X-axis) in a unit square (Figure C3b). The form of the ROC curve illustrates how HRs vary with FARs as one systematically lowers the threshold probability at which it is assumed that an event has been technically forecast to happen (i.e., from a 100% probability in the bottom left corner to a 0% probability at the top right corner).

The values of the geometrical area under the ROC curve (AUC-ROC) provide a summary measure of the discrimination ability across all probability thresholds. Perfect discrimination ability is obtained when only HRs grow, and FARs remain zero (Figure C3b). It is represented by an ROC curve that rises along the Y-axis from the bottom left corner of the unit square to the top left corner and moves straight to the top right corner. In this case, the AUC-ROC equals 1. If HRs and FARs grow at the

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

same rate, the forecasts may appear to lack discrimination ability, as they perform similarly to a climatological forecast or due to a limited number of issued forecasts exceeding the verifying rainfall thresholds. In this case, the ROC curve lies along the diagonal, and AUC-ROC equals 0.5.

How ROC curves and AUC-ROCs are computed can impact the interpretation of the forecasts' discrimination ability. The ROC curves will be built for 1% incremental decision thresholds that are meaningful for practical applications, and by 0.01% that are theoretically possible but would not be used in real-life applications. The ROC curves are built by straight segments joining successive points, and they are completed by joining the last meaningful point with a straight line in the top right corner of the unit square. For rare events (Figure C3b), the points of a ROC curve built with a 1% discretisation are more likely to cluster in the graph's bottom left corner, and completing the ROC with a straight line might give the impression that part of the ROC curve is missing (Casati et al., 2008). The area under the ROC curve (AUC-ROC) will be computed using a trapezoidal approximation by adding the areas of single trapeziums formed by the straight lines between consecutive points in the ROC curve (Bouallègue and Richardson, 2022).

Finally, the precision-recall (PR) curve is introduced to complement the discrimination analysis provided by ROC curves. PR curves are better suited than ROC curves for evaluating predictions trained on imbalanced datasets, as they focus on the minority class and are more sensitive to changes in false alarm rates when true positives are rare (Saito and Rehmsmeier, 2015; Juba and Le, 2019; Sofaer et al., 2019). The PR curve is built from the probabilistic contingency table in Figure C2b, mapping precision (in the y-axis) against recall (in the x-axis), computed from equations C4 and C5:

$$\text{precision} = \frac{\text{H}}{\text{H} + \text{FA}} \quad [\text{values between } 0 \text{ and } 1] \tag{C4}$$

$$\text{recall} = \frac{\text{H}}{\text{H} + \text{M}} \quad [\text{values between } 0 \text{ and } 1] \tag{C5}$$

It is worth noting that recall is better known in hydro-meteorology as hit rate (Equation C2). Moreover, PR curves are more commonly known in meteorology as "Performance Diagrams" and have been primarily applied to deterministic predictions (Taylor, 2001).

Code and data availability. The pre-processed training, test datasets, and forecast data in tabular format, as well as the trained XGBoost model underpinning all results presented in this paper, are archived on Zenodo under a CC BY-NC-ND 4.0 licence (DOI: `10.5281/zenodo.19140299`). We welcome academic collaboration and derivative use by request. The full code for model training, evaluation, and figure generation is available on the following GitHub repository: https://github.com/FatimaPillosu/probability_of_flash_flood (DOI: `https://doi.org/10.5281/zenodo.19144498`). The code for the retrieval of raw data and the pre-processing scripts that transform the raw input data into the processed point data tables are also included in the repository for full methodological transparency, but require access to the raw data described below. The impact reports from the Storm Event Database are available at https://www.ncei.noaa.gov/pub/data/swdi/

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

stormevents/csvfiles/. ERA5 reanalysis fields are publicly available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu); the ERA5 forecasts cannot be openly redistributed but are available from the corresponding author upon request.

Author contributions. F.M.P. designed the research, developed the model code, performed the simulations, analysed the data, produced the figures, and drafted the manuscript. M.C. supervised the machine learning model development. C.B. contributed to the interpretation of results. F.P., C.P., and H.L.C. supervised the study and secured the funding that supported this research. All authors contributed to the drafting and reviewing of the manuscript.

Competing interests. The authors have no competing interests to declare.

Acknowledgements. This research was conducted as part of the PhD project entitled "Outrunning flash floods: improving global medium-range forecasts for better preparedness" at the University of Reading, United Kingdom.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3292500.3330701, 2019.

Al-Rawas, G., Nikoo, M. R., and Al-Wardy, M.: A review on the prevention and control of flash flood hazards on a global scale: Early warning systems, vulnerability assessment, environmental, and public health burden, International Journal of Disaster Risk Reduction, 115, 105 024, https://doi.org/10.1016/j.ijdrr.2024.105024, 2024.

Altalhan, M., Algarni, A., and Turki-Hadj Alouane, M.: Imbalanced Data Problem in Machine Learning: A Review, IEEE Access, 13, 13 686–13 699, https://doi.org/10.1109/ACCESS.2025.3531662, 2025.

Aurelio, Y. S., de Almeida, G. M., de Castro, C. L., and Braga, A. P.: Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function, Neural Processing Letters, 50, 1937–1949, https://doi.org/10.1007/s11063-018-09977-1, 2019.

Ayodele, A.: A comparative study of ensemble learning techniques for imbalanced classification problems, World Journal of Advanced Research and Reviews, 19, 1633–1643, https://doi.org/10.30574/wjarr.2023.19.1.1202, 2023.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., Hurk, B. v. d., Hirschi, M., and Betts, A. K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, Journal of Hydrometeorology, 10, 623–643, https://doi.org/10.1175/2008JHM1068.1, 2009.

Barthold, F. E., Workoff, T. E., Cosgrove, B. A., Gourley, J. J., Novak, D. R., and Mahoney, K. M.: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment, Bulletin of the American Meteorological Society, 96, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1, 2015.

Bazo, J., Singh, R., Destrooper, M., and de Perez, E. C.: Pilot experiences in using seamless forecasts for early action: The "ready-set-go!" approach in the Red Cross, in: Sub-seasonal to seasonal prediction: The gap between weather and climate forecasting, pp. 387–398, Elsevier, https://doi.org/10.1016/B978-0-12-811714-9.00018-8, 2019.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J. P., Beljaars, A., and Bormann, N.: Representing equilibrium and nonequilibrium convection in large-scale models, Journal of the Atmospheric Sciences, 71, 734–753, https://doi.org/10.1175/JAS-D-13-0163.1, 2014.

Bertola, M., Blöschl, G., Bohac, M., Borga, M., Castellarin, A., Chirico, G. B., Claps, P., Dallan, E., Danilovich, I., Ganora, D., Gorbachova, L., Ledvinka, O., Mavrova-Guirguinova, M., Montanari, A., Ovcharuk, V., Viglione, A., Volpi, E., Arheimer, B., Aronica, G. T., Bonacci, O., Čanjevac, I., Csik, A., Frolova, N., Gnandt, B., Gribovszki, Z., Gül, A., Günther, K., Guse, B., Hannaford, J., Harrigan, S., Kireeva, M., Kohnová, S., Komma, J., Kriauciuniene, J., Kronvang, B., Lawrence, D., Lüdtke, S., Mediero, L., Merz, B., Molnar, P., Murphy, C., Oskoruš, D., Osuch, M., Parajka, J., Pfister, L., Radevski, I., Sauquet, E., Schröter, K., Šraj, M., Szolgay, J., Turner, S., Valent, P., Veijalainen, N., Ward, P. J., Willems, P., and Zivkovic, N.: Megafloods in Europe can be anticipated from observations in hydrologically similar catchments, Nature Geoscience, 16, 982–988, https://doi.org/10.1038/s41561-023-01300-5, 2023.

Bouallègue, Z. B. and Richardson, D. S.: On the ROC Area of Ensemble Forecasts for Rare Events, Weather and Forecasting, 37, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1, 2022.

Bucherie, A., Werner, M., Homberg, M. V. D., and Tembo, S.: Flash flood warnings in context: Combining local knowledge and large-scale hydro-meteorological patterns, Natural Hazards and Earth System Sciences, 22, 461–480, https://doi.org/10.5194/nhess-22-461-2022, 2022.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Bushra, S., Shakya, J., Cattoën, C., Fischer, S., and Pahlow, M.: CAMELS-NZ: hydrometeorological time series and landscape attributes for New Zealand, Earth System Science Data, 17, 5745–5760, https://doi.org/10.5194/essd-17-5745-2025, 2025.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorological Applications, 15, 3–18, https://doi.org/10.1002/met.52, 2008.

Cavaiola, M., Cassola, F., Sacchetti, D., Ferrari, F., and Mazzino, A.: Hybrid AI-enhanced lightning flash prediction in the medium-range forecast horizon, Nature Communications, 15, 1–15, https://doi.org/10.1038/s41467-024-44697-2, 2024.

Chen, Y.-M., Liu, C.-H., Shih, H.-J., Chang, C.-H., Chen, W.-B., Yu, Y.-C., Su, W.-R., and Lin, L.-Y.: An operational forecasting system for flash floods in mountainous areas in taiwan, Water, 11, 2100, https://doi.org/10.3390/w11102100, 2019.

Clark, R. A., Gourley, J. J., Flamig, Z. L., Hong, Y., and Clark, E.: CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products, Weather and Forecasting, 29, 377–392, https://doi.org/10.1175/WAF-D-12-00124.1, 2014.

Costache, R., Bao Pham, Q., Corodescu-Roșca, E., Cîmpianu, C., Hong, H., Thi Thuy Linh, N., Ming Fai, C., Najah Ahmed, A., Vojtek, M., Muhammed Pandhiani, S., Minea, G., Ciobotaru, N., Cristian Popa, M., Diaconu, D. C., and Thai Pham, B.: Using GIS, Remote Sensing, and Machine Learning to Highlight the Correlation between the Land-Use/Land-Cover Changes and Flash-Flood Potential, Remote Sensing, 12, 1422, https://doi.org/10.3390/rs12091422, 2020.

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., and Andréassian, V.: CAMELS-FR dataset: a large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking, Earth System Science Data, 17, 1461–1479, https://doi.org/10.5194/essd-17-1461-2025, 2025.

Dimitriadis, T., Gneiting, T., and Jordan, A. I.: Stable reliability diagrams for probabilistic classifiers, Proceedings of the National Academy of Sciences, 118, e2016191 118, https://doi.org/10.1073/pnas.2016191118, 2021.

Dordevic, M., Mutic, P., and Kim, H.: Flash Flood Guidance System: Response to one of the deadliest hazards, https://wmo.int/media/magazine-article/flash-flood-guidance-system-response-one-of-deadliest-hazards, 2020.

Doswell, C. A.: Severe Convective Storms—An Overview, in: Severe Convective Storms, edited by Doswell, C. A., pp. 1–26, American Meteorological Society, Boston, MA, 2001.

Dougherty, E. and Rasmussen, K. L.: Climatology of Flood-Producing Storms and Their Associated Rainfall Characteristics in the United States, Monthly Weather Review, 147, 3861–3877, https://doi.org/10.1175/MWR-D-19-0020.1, 2019.

Ebi, K. L., Vanos, J., Baldwin, J. W., Bell, J. E., Hondula, D. M., Errett, N. A., Hayes, K., Reid, C. E., Saha, S., Spector, J., and Berry, P.: Extreme weather and climate change: Population health and health system implications, Annual Review of Public Health, 42, 293–315, https://doi.org/10.1146/annurev-publhealth-012420-105026, 2021.

ECMWF: Part IV: Physical Processes, in: IFS Documentation CY41R2, IFS Documentation, ECMWF, https://doi.org/10.21957/tr5rv27xu, 2016.

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., Salamon, P., Brown, J. D., Hjerdt, N., Donnelly, C., Baugh, C. A., and Cloke, H. L.: Continental and global scale flood forecasting systems, Wiley Interdisciplinary Reviews: Water, 3, https://doi.org/10.1002/wat2.1137, 2016.

Flamig, Z. L. Z. L., Vergara, H., and Gourley, J. J.: The ensemble framework for flash flood forecasting (EF5) v1.2: Description and case study, Geoscientific Model Development, 13, 4943–4958, https://doi.org/10.5194/gmd-13-4943-2020, 2020.

Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H. X., Guerreiro, S., Haerter, J. O., Kendon, E. J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., and Zhang, X.: Anthropogenic intensification of short-duration rainfall extremes, Nature Reviews Earth & Environment, 2, 107–122, https://doi.org/10.1038/s43017-020-00128-6, 2021.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Gascón, E., Magnusson, L., Hewson, T., Rey, J., and Rodríguez, J.: Extreme precipitation in Spain's Valencia region, ECMWF Newsletter, 183, https://www.ecmwf.int/en/newsletter/183/news/extreme-precipitation-spains-valencia-region, 2025.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D., and Viglione, A.: A compilation of data on European flash floods, Journal of Hydrology, 367, https://doi.org/10.1016/j.jhydrol.2008.12.028, 2009.

Georgakakos, K. P., Modrick, T. M., Shamir, E., Campbell, R., Cheng, Z., Jubach, R., Sperfslage, J. A., Spencer, C. R., and Banks, R.: The flash flood guidance system implementation worldwide: a successful multidecadal research-to-operations effort, Bulletin of the American Meteorological Society, 103, E665–E679, https://doi.org/10.1175/bams-d-20-0241.1, 2022.

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P. E., Clark, R. A., Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J. M., Hong, Y., and Howard, K. W.: The FLASH project - improving the tools for flash flood monitoring and prediction across the united states, Bulletin of the American Meteorological Society, 98, https://doi.org/10.1175/BAMS-D-15-00247.1, 2017.

Grillakis, M. G., Koutroulis, A. G., Komma, J., Tsanis, I. K., Wagner, W., and Blöschl, G.: Initial soil moisture effects on flash flood generation – A comparison between basins of contrasting hydro-climatic conditions, Journal of Hydrology, 541, 206–217, https://doi.org/10.1016/j.jhydrol.2016.03.007, 2016.

Göber, M., Zsótér, E., and Richardson, D. S.: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification, Meteorological Applications, 15, 359–365, https://doi.org/10.1002/met.78, 2008.

Haiden, T., Janousek, M., Vitart, F., Prates, F., Maier-Gerber, M., Li, C. W. Y., and Chevallier, M.: Evaluation of ECMWF forecasts, ECMWF Technical Memoranda, 931, https://doi.org/10.21957/51e665e5c8, 2025.

Hancock, J., Johnson, J. M., and Khoshgoftaar, T. M.: A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data, in: 2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC), pp. 135–142, https://doi.org/10.1109/CIC56439.2022.00028, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P. d., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, https://doi.org/10.1002/qj.3803, 2020.

Hewson, T.: Capturing extreme rainfall events, ECMWF Newsletter, 2024.

Hewson, T. D. and Pillosu, F. M.: A low-cost post-processing technique improves weather forecasts around the world, Communications Earth & Environment, 2, 1–10, https://doi.org/10.1038/s43247-021-00185-9, 2021.

Hirabayashi, Y., Tanoue, M., Sasaki, O., Zhou, X., and Yamazaki, D.: Global exposure to flooding from the new CMIP6 climate model projections, Scientific Reports, 11, https://doi.org/10.1038/s41598-021-83279-w, 2021.

Ibarreche, J., Aquino, R., Edwards, R. M., Rangel, V., Pérez, I., Martínez, M., Castellanos, E., Álvarez, E., Jimenez, S., Rentería, R., Edwards, A., and Álvarez, O.: Flash Flood Early Warning System in Colima, Mexico, Sensors, 20, 5231, https://doi.org/10.3390/s20185231, 2020.

IPCC: Climate change 2023: Synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change [core writing team, H. Lee and J. Romero (eds.)], IPCC, Geneva, Switzerland, https://doi.org/10.59327/IPCC/AR6-9789291691647, 2023.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Iqbal, J., Bux, H., and Sahitia, S.: Health Consequences of Natural Disasters: An Overview of Recent Literature on Floods, Pakistan Journal of Public Health, 13, 192–199, https://doi.org/10.32413/pjph.v13i4.1287, 2023.

Javelle, P., Organde, D., Demargne, J., Saint-Martin, C., de Saint-Aubin, C., Garandeau, L., and Janet, B.: Setting up a French national flash flood warning system for ungauged catchments based on the AIGA method, vol. 7, pp. 1–11, https://doi.org/10.1051/e3sconf/20160718010, 2016.

Jimenez, D. A., Meneses, J. E., Solha, P. H. B., Avila-Diaz, A., Quesada, B., Melo Brentan, B., and Ferreira Rodrigues, A.: CAMELS-COL: A Large-Sample Hydrometeorological Dataset for Colombia, Earth System Science Data Discussions, pp. 1–38, https://doi.org/10.5194/essd-2025-200, 2025.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley & Sons, second edn., 2012.

Juba, B. and Le, H. S.: Precision-Recall versus Accuracy and the Role of Large Data Sets, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 4039–4048, https://doi.org/10.1609/aaai.v33i01.33014039, 2019.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Scientific Data, 10, 1–11, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS opinions: Never train a long short-term memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Lavers, D. A., Harrigan, S., and Prudhomme, C.: Precipitation biases in the ECMWF integrated forecasting system, Journal of Hydrometeorology, 22, 1315–1334, https://doi.org/10.1175/jhm-d-20-0308.1, 2021.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J.: Evaluation of ERA5 precipitation for climate monitoring, Quarterly Journal of the Royal Meteorological Society, 148, 2776–2788, https://doi.org/10.1002/QJ.4351, 2022.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N.: A survey on addressing high-class imbalance in big data, Journal of Big Data, 5, 42, https://doi.org/10.1186/s40537-018-0151-6, 2018.

Liu, C., Guo, L., Ye, L., Zhang, S., Zhao, Y., and Song, T.: A review of advances in China's flash flood early-warning system, Natural Hazards, 92, 619–634, https://doi.org/10.1007/s11069-018-3173-7, 2018.

Liu, J., Koch, J., Stisen, S., Troldborg, L., Højberg, A. L., Thodsen, H., Hansen, M. F. T., and Schneider, R. J. M.: CAMELS-DK: hydrometeorological time series and landscape attributes for 3330 Danish catchments with streamflow observations from 304 gauged stations, Earth System Science Data, 17, 1551–1572, https://doi.org/10.5194/essd-17-1551-2025, 2025.

Luo, L., Wang, Y., Li, Q., Li, M., Wang, J., Zhao, G., and Ma, M.: Exploration of the spatiotemporal characteristics and triggering factors of flash flood in China, Ecological Indicators, 176, 113 698, https://doi.org/10.1016/j.ecolind.2025.113698, 2025.

López, V., Fernández, A., and Herrera, F.: On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed, Information Sciences, 257, 1–13, https://doi.org/10.1016/j.ins.2013.09.038, 2014.

Marjerison, R. D., Walter, M. T., Sullivan, P. J., and Colucci, S. J.: Does Population Affect the Location of Flash Flood Reports?, Journal of Applied Meteorology and Climatology, 55, 1953–1963, https://doi.org/10.1175/JAMC-D-15-0329.1, 2016.

Matthews, G., Baugh, C., Barnard, C., Carton De Wiart, C., Colonese, J., Grimaldi, S., Ham, D., Hansford, E., Harrigan, S., Heiselberg, S., Hooker, H., Hossain, S., Mazzetti, C., Milano, L., Moschini, F., O'Regan, K., Pappenberger, F., Pfister, D., Rajbhandari, R. M., Salamon, P., Ramos, A., Shelton, K., Stephens, E., Tasev, D., Turner, M., van den Homberg, M., Wittig, J., Zsótér, E., and Prudhomme, C.: Chapter 15 - on the operational implementation of the global flood awareness system (GloFAS), in: Flood forecasting (second edition), edited

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

by Adams, T. E., Gangodagamage, C., and Pagano, T. C., pp. 299–350, Academic Press, second edition edn., ISBN 978-0-443-14009-9, https://doi.org/https://doi.org/10.1016/B978-0-443-14009-9.00014-6, 2025.

Maybee, B., Birch, C. E., Böing, S. J., Willis, T., Speight, L., Porson, A. N., Pilling, C., Shelton, K. L., and Trigg, M. A.: FOREWARNS: development and multifaceted verification of enhanced regional-scale surface water flood forecasts, Natural Hazards and Earth System Sciences, 24, 1415–1436, https://doi.org/10.5194/nhess-24-1415-2024, 2024.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Nijzink, J., Loritz, R., Gourdol, L., Zoccatelli, D., Iffly, J. F., and Pfister, L.: CAMELS-LUX: Highly Resolved Hydro-Meteorological and Atmospheric Data for Physiographically Characterized Catchments around Luxembourg, Earth System Science Data Discussions, pp. 1–34, https://doi.org/10.5194/essd-2024-482, 2025.

NWS: NOAA's National Weather Service - Glossary, https://forecast.weather.gov/glossary.php?word=flash+flood, 2025.

Panwar, V. and Sen, S.: Disaster Damage Records of EM-DAT and DesInventar: A Systematic Comparison, Economics of Disasters and Climate Change, 4, 295–317, https://doi.org/10.1007/s41885-019-00052-0, 2020.

Pillosu, F.: Outrunning flash floods: improving global medium-range forecasts for better preparedness, PhD, University of Reading, Reading, UK, _, 2026.

Pillosu, F. M., Bucherie, A., Kruczkiewicz, A., Haiden, T., Baugh, C., Hultquist, C., Vergara, H., Pappenberger, F., Stephens, E., Prudhomme, C., and Cloke, H.: Can global rainfall forecasts identify areas at flash flood risk? Proof of concept for Ecuador, ECMWF Technical Memoranda, 917, 1–37, https://doi.org/10.21957/8e2dd559f0, 2024.

Pillosu, F. M., Hewson, T. D., Prudhomme, C., Gascòn, E., Vuckovic, M., Stephens, E., and Cloke, H.: Bridging the scale gap: enhancing point-scale rainfall estimates through the post-processing of ERA5, Unpublished, pp. 1–15, 2025.

Ranglani, H.: Empirical Analysis Of The Bias-Variance Tradeoff Across Machine Learning Models, Machine Learning and Applications: An International Journal (MLAIJ), 11, https://doi.org/10.2139/ssrn.5086450, 2024.

Raynaud, D., Thielen, J., Salamon, P., Burek, P., Anquetin, S., and Alfieri, L.: A dynamic runoff co-efficient to improve flash flood early warning in Europe: Evaluation on the 2013 central European floods in Germany, Meteorological Applications, 22, https://doi.org/10.1002/met.1469, 2015.

Rezaei-Dastjerdehei, M. R., Mijani, A., and Fatemizadeh, E.: Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function, in: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), pp. 333–338, https://doi.org/10.1109/ICBME51989.2020.9319440, 2020.

Robbins, J. C. and Titley, H. A.: Evaluating high-impact precipitation forecasts from the Met Office Global Hazard Map (GHM) using a global impact database, Meteorological Applications, 25, 548–560, https://doi.org/10.1002/met.1720, 2018.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R.: The Shapley Value in Machine Learning, in: Proceedings of the 31st International Joint Conference on Artifical Intelligence, IJCAI-ECAI 2022, pp. 5572–5579, International Joint Conferences on Artificial Intelligence Organization, https://doi.org/10.24963/ijcai.2022/778, 2022.

Sadkou, S., Artigue, G., Fréalle, N., Ayral, P.-A., Pistre, S., Sauvagnargues, S., and Johannet, A.: A review of flash-floods management: From hydrological modeling to crisis management, Journal of Flood Risk Management, 17, e12 999, https://doi.org/10.1111/jfr3.12999, 2024.

Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., Hong, Y., and Giroud, M.: Mapping Flash Flood Severity in the United States, Journal of Hydrometeorology, 18, 397–411, https://doi.org/10.1175/JHM-D-16-0082.1, 2017.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Saito, T. and Rehmsmeier, M.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, PLOS ONE, 10, e0118 432, https://doi.org/10.1371/journal.pone.0118432, 2015.

Santos, L., Satolo, L., Oyarzabal, R., Escobar-Silva, E., Diniz, M., Negri, R., Lima, G., Stephany, S., Soares, J., Duque, J., Saraiva-Filho, F., and Bacelar, L.: Machine Learning-based Hydrological Models for Flash Floods: A Systematic Literature Review, Smart Construction and Sustainable Cities, 3, https://doi.org/10.1007/s44268-025-00071-9, 2025.

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., and Patil, K. R.: Overview of leakage scenarios in supervised machine learning, Journal of Big Data, 12, 135, https://doi.org/10.1186/s40537-025-01193-8, 2025.

Schumacher, R. S.: Heavy rainfall and flash flooding, in: Oxford research encyclopedia of natural hazard science, pp. 1–42, Oxford University Press, https://doi.org/10.1093/acrefore/9780199389407.013.132, 2017.

Shwartz-Ziv, R. and Armon, A.: Tabular data: Deep learning is not all you need, Information Fusion, 81, 84–90, https://doi.org/10.1016/j.inffus.2021.11.011, 2022.

Singh, N. K., Emanuel, R. E., McGlynn, B. L., and Miniat, C. F.: Soil moisture responses to rainfall: Implications for runoff generation, Water Resources Research, 57, https://doi.org/10.1029/2020WR028827, 2021.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S.: The area under the precision-recall curve as a performance metric for rare binary events, Methods in Ecology and Evolution, 10, 565–577, https://doi.org/10.1111/2041-210X.13140, 2019.

Speight, L. J., Cranston, M. D., White, C. J., and Kelly, L.: Operational and emerging capabilities for surface water flood forecasting, Wiley Interdisciplinary Reviews: Water, 8, e1517, https://doi.org/10.1002/wat2.1517, 2021.

Strumbelj, E. and Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory, J. Mach. Learn. Res., 11, 1–18, 2010.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Tsonevsky, I., Doswell, C. A., and Brooks, H. E.: Early warnings of severe convection using the ECMWF extreme forecast index, Weather and Forecasting, 33, https://doi.org/10.1175/WAF-D-18-0030.1, 2018.

UN: Early Warnings for All, https://www.un.org/en/climatechange/early-warnings-for-all, 2022.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, fourth edn., ISBN 978-0-12-815823-4, https://doi.org/10.1016/C2017-0-03921-6, 2020.

WMO, W. M. O.: State of the climate 2024, https://wmo.int/publication-series/state-of-global-climate-2024, 2025.

Ying, X.: An Overview of Overfitting and its Solutions, Journal of Physics: Conference Series, 1168, 022 022, 2019.

Zanchetta, A. D. L. and Coulibaly, P.: Recent advances in real-time pluvial flash flood forecasting, Water (Switzerland), 12, https://doi.org/10.3390/w12020570, 2020.

Zhai, X., Guo, L., Liu, R., and Zhang, Y.: Rainfall threshold determination for flash flood warning in mountainous catchments with consideration of antecedent soil moisture and rainfall pattern, Natural Hazards, 94, https://doi.org/10.1007/s11069-018-3404-y, 2018.

Zhang, Y., Li, Z., Xu, H., Ge, W., Qian, H., Li, J., Sun, H., Zhang, H., and Jiao, Y.: Impact of floods on the environment: A review of indicators, influencing factors, and evaluation methods, Science of The Total Environment, 951, 175 683, https://doi.org/10.1016/j.scitotenv.2024.175683, 2024.

Zhao, C., Liu, J., and Parilina, E.: The Shapley Value Contribution to Explainable Artificial Intelligence: A Comprehensive Survey, Dynamic Games and Applications, https://doi.org/10.1007/s13235-025-00670-2, 2025a.

https://doi.org/10.5194/egusphere-2026-1591
Preprint. Discussion started: 8 April 2026

Zhao, Y., Wu, X., Guo, L., Qin, G., Li, X., and Li, H.: Evaluation of rainfall-threshold methods for flash flood warnings based on soil moisture conditions, Natural Hazards, https://doi.org/10.1007/s11069-025-07272-6, 2025b.

Zhao, Y., Wu, X., Zhang, W., Lan, P., Qin, G., Li, X., and Li, H.: A deep learning-based probabilistic approach to flash flood warnings in mountainous catchments, Journal of Hydrology, 652, 132 677, https://doi.org/10.1016/j.jhydrol.2025.132677, 2025c.

Žagar, N.: A global perspective of the limits of prediction skill of NWP models, Tellus A: Dynamic Meteorology and Oceanography, 69, 1317 573, https://doi.org/10.1080/16000870.2017.1317573, _eprint: https://doi.org/10.1080/16000870.2017.1317573, 2017.