https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Mapping uncertainty in flood impacts: a multi-source assessment and catalogue of simulated, remotely sensed and reported floods in Europe

Lorenzo Scarpellini[1,2], Andrea Ficchì[1], Claudia D'Angelo[3], Andrea Betterle[3], Peter Salamon[3], and Andrea Castelletti[1,4]

[1]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy
[2]University School for Advanced Studies of Pavia, Piazza della Vittoria, 15, 27100, Pavia, Italy
[3]European Commission, Joint European Research Centre (JRC), Via Enrico Fermi 2749, 21027, Ispra, Italy
[4]Euro-Mediterranean Center on Climate Change (CMCC), Via Savona 16, 20144, Milano, Italy

Correspondence: Lorenzo Scarpellini (lorenzo.scarpellini@polimi.it) and Andrea Castelletti (andrea.castelletti@polimi.it)

Abstract. Flooding is Europe's most costly natural hazard, with riverine floods alone accounting for over a third of disaster related damages, which have increased nearly tenfold since the 1990s. Despite advances in large-scale hydrological modeling and flood hazard mapping, continental-scale flood risk assessments remain highly uncertain, also due to simplified representations of flood protections and limitations in observational datasets and reported impacts. In this study, we assess how flood protection assumptions and data sources influence flood impact and risk estimates across Europe. We compile and provide a catalogue of simulated flood events based on the European Flood Awareness System (EFAS, version 5.0) modeling chain and quantify differences across three flood protection scenarios. Simulated events are then systematically compared with satellite-derived flood extents and impacts from the Copernicus Global Flood Monitoring (GFM) system, as well as with reported impacts from the HANZE database, using an automated event-matching procedure. Results indicate that flood protections are a major source of uncertainty, strongly affecting simulated flood frequency, spatial distribution, and damages. Moreover, substantial inconsistencies are found between simulations, observations, and reported data. Simulations tend to overestimate flood extents and damages, especially for large events, while satellite observations underestimate small and short-lived floods, particularly in urban environments. Reported datasets capture only a subset of impactful events and are biased toward populated regions, resulting in limited agreement across datasets and in differing regional impact patterns. Overall, results show that no single dataset can be currently considered a ground truth for flood impacts at continental scale, highlighting the need for multi-dataset approaches in large-scale flood risk assessment. These findings contribute to a more robust interpretation of flood risk estimates and support the development of more reliable and uncertainty-aware multi-dataset approaches for continental-scale flood risk assessments.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

1 Introduction

Flooding is currently Europe's most costly natural hazard, with riverine floods alone accounting for more than one third of all disaster related damages. Moreover, flood related economic losses are rising and are now almost ten times higher than in the 1990s (European Environment Agency, 2025; Joint Research Centre of the European Commission, 2026). For these reasons, considerable effort has been devoted to improving the resolution, quality, and reliability of hydrological and flood hazard models, as well as to developing new quantitative approaches for continental-scale flood risk assessments. In fact, several large-scale hydrological models and flood hazard maps at continental or global scale are now available, both publicly and commercially (Aerts et al., 2020; Bernhofen et al., 2018), allowing for fast and more accurate decision-making applications at previously unattainable spatial scale and resolution.

Despite these advances, large-scale flood hazard and loss models are affected by several sources of uncertainty (e.g., Pianosi et al., 2026), ranging from meteorological forcings, hydrological and hydro-dynamic model structure to parameter uncertainties and modeling simplifications, such as the representation of flood protections. Flood protections are in fact often integrated at the end of the workflow, by discarding floods simulated at a specific location if the corresponding return period is lower than the estimated or assumed protection level (e.g., Dottori et al., 2023; Riedel et al., 2024; Toosi et al., 2020; Coccia et al., 2023; Aerts et al., 2020). Currently the FLOPROS dataset (Scussolini et al., 2016) is the most widely used for this purpose, thanks to its level of detail and its global coverage, but recently new regional datasets have also started to emerge. One of these was presented in Paprotny et al. (2025) for Europe (hereafter referred to as $P25$), which offers an alternative to FLOPROS with much lower estimated protection levels. In parallel, several studies have sought to estimate river levee heights to enable direct integration into hydraulic and hydrodynamic models (Boulange et al., 2025; Khanh et al., 2025; Zhao et al., 2025), whose development holds promise for a much better representation of flood risk in the future. However, all currently available flood protection datasets rely on simplifying assumptions and incomplete information, and none of them can be considered a ground truth representation of flood protection levels. Another common approximation is to assign a single return period (RP) to a flood event across the affected river network, a practical but simplistic assumption that introduces additional uncertainty into the analysis. In reality, during the same flood event, different locations may experience water depths associated with different return periods, particularly in areas with multiple rivers and confluences.

Given these uncertainties, the validation and systematic assessment of large-scale flood hazard model outputs is essential. Traditionally, flood hazard models have been validated in three different ways: (i) through expert-based judgment, (ii) by inter-comparisons with other models, and (iii) by systematic evaluation against observed data (Molinari et al., 2019). While expert-based judgment is not feasible at continental scale and model inter-comparisons (and comparisons with reference maps) have already been largely explored in past studies (e.g. Trigg et al. (2016); Bernhofen et al. (2018); Aerts et al. (2020); Risling et al. (2024); Dottori et al. (2016)), quantitative evaluations against observations have until now been limited to a small number of selected case studies and flood events (Bernhofen et al., 2018; Choné et al., 2021; Wing et al., 2020; Risling et al., 2024).

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

As observed data, most past assessment studies have relied on dataset deriving from optical satellite images, which are however cursed by several limitations including being able to observe floods only during daytime and over cloud-free areas, a strong limitation for flood delineation. In particular, such images tend to produce a significant amount of false alarms in dark surfaces, in addition to being known to under perform when sediments are suspended in the water (Portalés-Julià et al., 2023). In addition, optical satellites with high revisit times also have low or medium resolution (e.g., MODIS). Some of these limitations can be overcome with Synthetic Aperture Radar (SAR) satellites, which can detect flooded areas day and night and regardless of cloud cover. At the same time though, SAR is still limited by potentially lower revisit times and by the similarity of the back scattering of different objects and land covers to water, with the consequent potential identification of false alarms (Risling et al., 2024; Betterle and Salamon, 2025). Currently the best publicly available source of SAR data is the Sentinel-1 mission, part of the European Copernicus program. Sentinel-1 data have been extensively used in recent years for the identification of floods (Matgen et al., 2020; Salamon et al., 2021; Twele et al., 2016; Bauer-Marschallinger et al., 2022). Their increasing adoption culminated in the recent release of the Global Flood Monitoring system (GFM) of the Copernicus Emergency Management Service (CEMS). This system automatically generates near real-time flood delineation maps from Sentinel-1 acquisitions and freely distributes them through an online portal. Until recently flood extent was the only information available through this system, but Betterle and Salamon (2025) used topographic information to extract water depth information from flood extent thanks to a novel algorithm (Betterle and Salamon, 2024), allowing to estimate pseudo-observed flood damages.

Given these recent developments, this study aims to systematically assess the role of flood protection and data sources assumptions in large-scale flood impact assessments across Europe. Specifically, we first compile a new dataset of simulated flood events based on the European Flood Awareness System (EFAS, part of the Copernicus Emergency Management Service). EFAS is based on a continental-scale hydrological modeling chain, using the Open Source LISFLOOD hydrological model. We then quantify how different representations of flood protection levels affect simulated flood frequency, spatial distribution, and associated damages, by considering the FLOPROS and P25 datasets, as well as a scenario without flood protections.

Building on this, we perform a large-scale comparison between simulated floods, satellite-based observations, and reported flood impacts. In particular, for the first time at the European scale, simulated flood events are systematically matched with observed flood extents and pseudo-observed impacts derived from the Copernicus Global Flood Monitoring (GFM) system, without pre-selecting specific case studies. The event-based comparison is complemented by the inclusion of reported impacts from the Hanze dataset (Paprotny et al., 2024), enabling a three-way assessment of simulated, observed, and reported floods within a consistent analytical framework. The approach to the simulation of floods closely follows the methodology of Dottori et al. (2017), accounting for spatial variability in flood magnitudes within events and thereby avoiding the simplifying assumption of a single return period per flood.

Rather than validating a single dataset against a single presumed ground truth, the aim of this study is to quantify agreements, discrepancies, and systematic biases between flood simulations, satellite-derived observations, and reported impacts. This approach allows for a more comprehensive understanding of the strengths and limitations of each data source, and of how these propagate into flood impact estimates.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

The contributions of this work are threefold. First, we provide a systematic quantification of the sensitivity of large-scale flood impact estimates to flood protection assumptions, demonstrating their primary role in shaping flood hazard and risk. Second, we present the first continental-scale event-based comparison between simulated floods, satellite observations, and reported impacts, highlighting the similarity and inconsistencies between datasets. Third, we identify key sources of bias in both modeling and observational approaches and provide recommendations for improving large-scale flood risk assessments. Overall, this study contributes to a more robust interpretation of flood risk estimates and supports the development of more reliable and informed decision-making frameworks.

Finally, we also provide a publicly available catalogue of simulated flood events under different flood protection settings, together with a dataset of simulated and observed flood extents for events identified across multiple datasets, including those consistently detected in all three sources, to support further research and policy efforts.

2 Methods

The overall workflow adopted in this study, illustrated in figure 1, is divided into two main parts.

The first part involves the production and evaluation of simulated historical flood events. River discharge data are first converted into return periods (RP), after which flood peaks are identified. These peaks are then clustered in space and time in order to distinguish individual flood events. For each event, the flooded extent and associated impacts are estimated accounting for flood protection levels. The resulting simulations under different protection assumptions are subsequently compared with one another and with reported impact datasets. The second part incorporates satellite-based observations of flooded areas and water depths. After estimating the corresponding impacts from these observations, simulated flood events are matched with observed and reported events. At the same time, events present in only one datasets are identified, enabling the calculation of metrics that quantify agreement between simulated, observed and reported flood products.

The following sections describe each step of the workflow in detail, beginning with the description of the datasets that were used and the pre-processing procedures applied to them.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

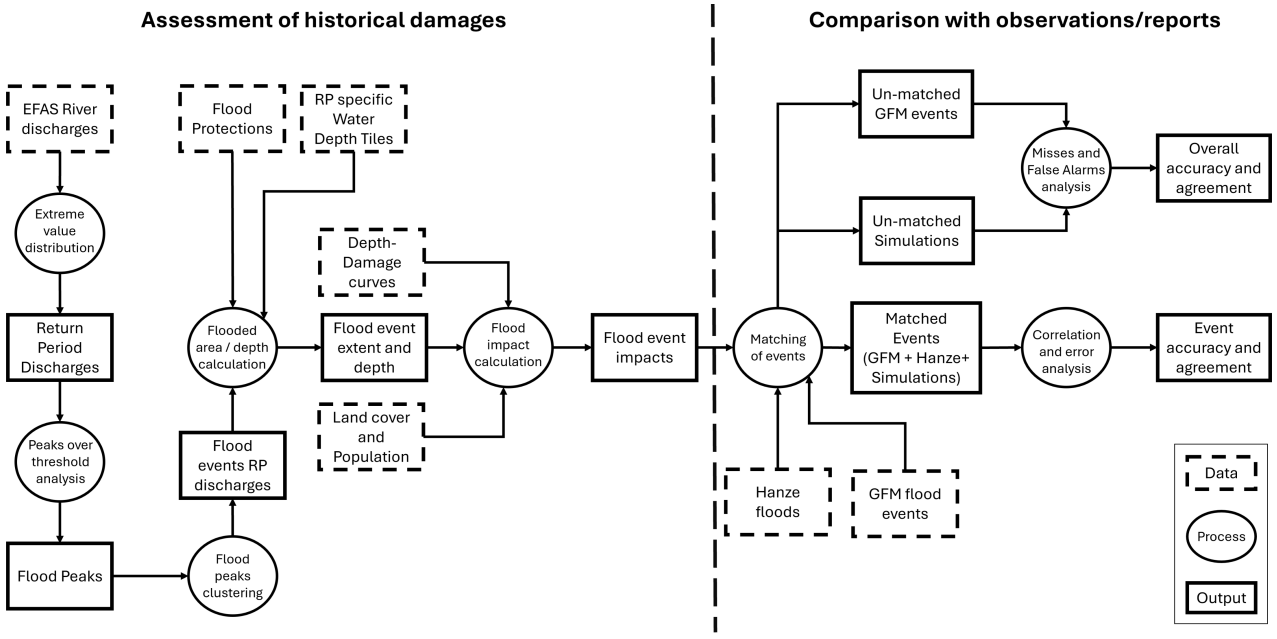

Figure 1. Workflow, with data, processes, and outputs

2.1 Data and Pre-processing

Administrative Units and Hydrological Basins

The European Nomenclature of Territorial Units for Statistics (NUTS) provides a standardized framework for spatial and statistical comparisons across European regions and countries. NUTS boundaries are updated approximately every four years, with the most recent version being NUTS 2024. However, in this study the NUTS 2016 classification (Eurostat, 2018), was used, as the study area includes the United Kingdom (data available on eurostat's website). Analyses were conducted at NUTS levels 0, 2, and 3.

Hydrological basins were downloaded from the FAO hydrographic dataset (FAO, 2021). The basins correspond approximately to level-2 catchments, with level-1 catchments encompassing the entire drainage areas of major rivers that flow into seas or oceans and level-2 catchments corresponding to the direct tributaries to main river of level-1 catchments.

River Discharge Data and Extreme Value Distribution

Historical river discharge simulations were extracted from CEMS-EFAS reanalysis, version 5.0 (Mazzetti et al., 2023), covering the period 1992–2024. These data were produced by forcing the operational version of the hydrological model LISFLOOD (Van Der Knijff et al., 2010; Burek et al., 2013) (documentation available online) with gridded observational data of precipitation and temperature, and are provided at a spatial resolution of around $1.5km$ $(0,016$ degrees in WGS84) and at temporal resolution of 6 hours. Data were then clipped to the EFAS river grid, with a minimum upstream area of $150km^2$, and later resampled to

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

daily mean discharge. Finally, a unique ID was assigned to each pixel. The discharge values were then converted from volumes to return periods, using the parameters of the Gumbel extreme value distribution provided in the EFAS Auxiliary Data.

Flood Hazard Maps and Tiles

Flood hazard maps provide the relationship between discharge return periods and the resulting flooded extent and water depth. These maps are typically derived from hydraulic simulations that represent both river channels and floodplain processes and are commonly distributed as composite hazard maps covering large spatial domains. In this study, flood hazard maps produced by the Joint Research Centre (JRC) of the European Commission were used (Baugh et al., 2024; Dottori et al., 2022; Alfieri et al., 2014). These maps were originally derived from simulations of the LISFLOOD-FP hydraulic model (Bates et al., 2010), and were downloaded from the JRC data portal. Flood hazard maps are available for nine return periods: 10, 20, 30, 50, 75, 100, 200, and 500 years. The 500-year return period map was used to represent the maximum potential flood extent across Europe. Consequently to each floodplain pixel within the 500-year return period flood hazard map was assigned a unique id ($pdam_{id}$). For the simulation of individual flood events, flood hazard tiles were used instead of composite maps. These tiles represent the underlying datasets from which the composite hazard maps are generated. Each tile corresponds to a specific river network pixel and return period, providing the full potential flooded area associated to a specific river location as well as the corresponding water depth distribution. The tiles, therefore, also define the spatial influence area of each river pixel for every given return period. Flood hazard tiles were produced approximately every two river pixels in the EFAS network and are not publicly available, but may be obtained upon motivated request.

Both the composite hazard maps and hazard tiles were originally produced at 90 m spatial resolution in the WGS84 coordinate reference system (EPSG:4326). For consistency with other datasets used in this study, they were reprojected to the Equi7-Europe grid (EPSG:27704) and resampled to a spatial resolution of 100 m using nearest-neighbor interpolation.

River Flood Protections

Two river flood protection datasets were considered in this study. The first is FLOPROS (Scussolini et al., 2016), a global database of flood protection standards expressed in terms of protection return periods. FLOPROS provides four layers: policy, design, modeled, and merged. In this study, only the merged layer was used, as it integrates information from the other layers to represent the best available estimate of protection levels. Since FLOPROS is not directly aligned with the NUTS classification, protection values were reassigned to each NUTS3 region based on the dominant overlap between the NUTS region and the FLOPROS spatial layer.

The second dataset is the P25 flood protection dataset, developed in Paprotny et al. (2025). This dataset provides protection levels directly at NUTS3 resolution and includes yearly values from 1950 to 2020. To ensure consistency between the two protection datasets and avoid temporal variability in protection levels, the mean protection value for the period 2015–2020 was calculated for each NUTS3 region and used throughout the analysis. More information regarding flood protections are available in Section A1.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Population and Land Cover

Population exposure was derived from the 2020 population grid of the Global Human Settlement Layer (GHSL, Schiavina et al. (2023)), the dataset was resampled to a spatial resolution of 100 m. Population values were then rounded to the nearest integer, and values between 0 and 1 were set equal to 1 to ensure representation of sparsely populated cells.

Land-cover information was obtained from the LUISA 2018 dataset (Pigaiani and Batista E Silva, 2021), which provides high spatial (100 m) and thematic (46 classes) resolution. The original LUISA land-cover classes were aggregated into five broader categories for this analysis: Residential, Industrial and Business, Agriculture, Transport, and Other (for more information, see Table A1).

Depth-Damage Curves

Flood damages were estimated using depth–damage curves based on Huizinga et al. (2017), updated to match the land-cover classification of the LUISA dataset. These curves express fractions of a maximum damage values as a function of water depth, reaching saturation at a depth of 6 m. Maximum damage values, originally provided at NUTS2 resolution, were downscaled to NUTS3 resolution using regional GDP statistics obtained from Eurostat (Eurostat, 2025).

Remote Sensed Flood Extents and Depths

Satellite-derived flood extents and corresponding water depth estimates for the period 2015–2024 were obtained from the JRC data portal (Betterle and Salamon, 2025). To ensure consistency with the simulation framework, each image was first clipped to the extent of the 500-year return period composite flood hazard map. This step ensured that areas outside the modeled floodplain were excluded, as these areas could not be represented in the riverine flood simulations and may correspond to other types of floods. The original maps, provided at 20 m resolution, were resampled to 100 m using average aggregation, aligning both the spatial resolution and coordinate reference system with the other datasets used in the analysis.

Reported Impacts

Reported flood impacts were obtained from the Hanze database (Paprotny et al., 2024), which contains information on flood events and associated impacts across Europe since 1870. Although impact data are not available for all events, the dataset represents one of the most comprehensive publicly available sources of historical flood impact information. When reported impacts were available for events affecting multiple NUTS3 regions, the total impacts were evenly distributed among all affected regions. While this assumption represents an obvious simplification, it was necessary to enable consistent comparisons with simulated and satellite-derived flood impact estimates.

2.2 Discharge Return Periods and Peaks

Discharge return periods form the basis of the flood risk assessment, as flooded areas are identified by selecting the pre-simulated inundation tiles corresponding to river pixels with a discharge return period (RP) exceeding 10 years.

Flood peaks were first identified for each river pixel using the $findpeaks$ function from the $scipy$ Python package. A minimum peak height corresponding to a return period of 10 years and a minimum temporal distance of 90 days between peaks were

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

imposed. The former constraint was necessary as a RP of 10 years is the minimum necessary in order to be able to use the pre-simulated hazard maps, the latter instead ensures that the likelihood of detecting multiple peaks associated with the same flood event is reduced or null.

Although multiple hydrological floods can occur within a three months period, the objective of this study is not to identify individual hydrological flood waves but rather to estimate their economic impacts. From an impact perspective, two floods occurring in rapid succession within the same area would not necessarily produce twice the damage of a single event, as recovery processes must be considered. Empirical evidence suggests that recovery times following extreme events can in fact be substantial. For instance, a recent study estimated that restoring 60–70 % of the economic value of assets damaged by extreme weather events may take up to one year (Platt et al., 2025). Other analyses have reported a slow-down in economic growth persisting for up to five months after flooding (Collalti, 2024) while the mean recovery time after a 5-year flood has been estimated at approximately 77 days (Johnson et al., 2024). The 90 day separation threshold, therefore, provides a pragmatic compromise between hydrological realism and impact-based event representation. For more information regarding the impact of this threshold, refer to section B5.

After identifying discharge peaks for each river pixel, nine return period classes were defined, ranging from 10 years to more than 500 years. Each class was assigned a discrete probability of occurrence based on the inverse of the return period (Equation 1). For example, the RP range 10–20 years was assigned a probability of 0.05 (1/10 - 1/20), while the ranges 40–50 years and 100–200 years both correspond to probabilities of 0.005. For events exceeding a 500-year return period, a probability of 0.002 was assigned.

$$\tilde{p}_i = \frac{1}{RP_i} - \frac{1}{RP_{i+1}}, \qquad RP_{K+1} = \infty \tag{1}$$

with i representing the return periods classes, the maximum being $K = 9$, representing a RP of 500 years.

These probabilities were later used with the composite flood hazard maps in order to calculate the expected annual damages.

2.3 Flood peaks clustering

The identified discharge peaks were subsequently aggregated into flood events based on their spatial and temporal proximity. Although several clustering algorithms exist for this purpose, such as DBSCAN and its spatio-temporal extensions (Ester et al., 1996; Birant and Kut, 2007)) a custom clustering algorithm was developed to allow greater control over the clustering parameters. Starting from the peak with the earliest occurrence date and proceeding chronologically, the algorithm runs considering at each iteration only peaks not already assigned to any event. The clustering procedure followed the steps below:

1. Flood peaks occurring within 90 days of the first peak and within a radius of 50 km were identified and assigned to the same cluster.
2. The convex hull encompassing all identified peaks (including the initial peak) was computed.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

3. The convex hull was iteratively expanded by searching for additional peaks occurring within 90 days of the initial peak and within a distance of 50 km from the convex hull boundary.
4. The iterative expansion continued until no further peaks were identified or until a maximum of 50 iterations was reached.
5. The next unassigned peak was then selected and the procedure repeated.

This procedure has two main advantages: it avoids random initialization and allows clusters of arbitrary size and shape. A schematic representation of the clustering procedure is provided in Figure 2, while for a more detailed description of the algorithm, refer to Section A1.

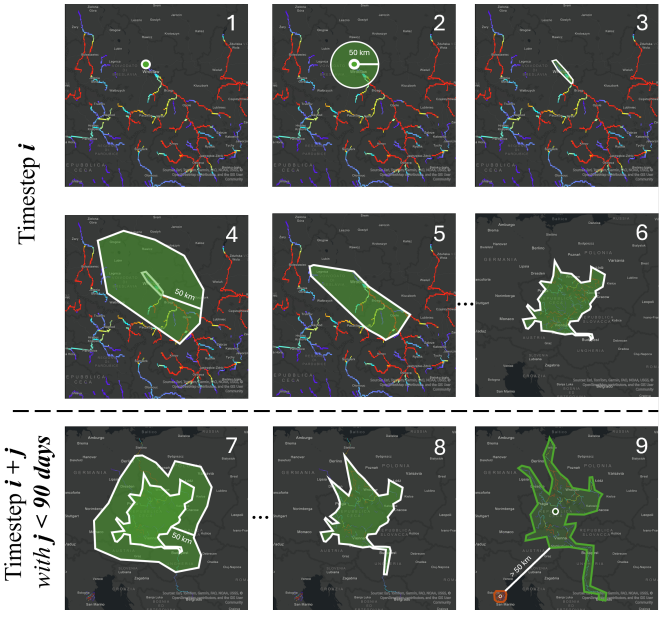

Figure 2. Visual representation of the peak clustering algorithm.

The output of this process is a catalogue of flood events, each characterized by start and end dates (corresponding to the dates of the first and last peak respectively), associated peak discharges, and a geographic location defined as the convex hull enclosing the discharge peaks.

2.4 Flood extent and impacts calculation

Once flood events and their associated discharge peaks were identified, the flooded extents and impacts were calculated. The tiles corresponding to each river pixel of the clustered flood event were then combined to derive the maximum flooded extent and maximum water depth for the event, following the methodology described in Dottori et al. (2017). A key difference introduced in this study concerns the treatment of flood protection levels. In (Dottori et al., 2017), only hazard tiles associated to river pixels whose discharge exceeded the protection return period of the respective region are considered, so that all other

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

tiles are entirely excluded. In the present study, hazard tiles are on the contrary always considered, but filtering is applied at the floodplain-pixel level: for each floodplain pixel potentially affected during an event, first only the flood depth of tiles with RP exceeding the protection level of the NUTS3 region in which the floodplain pixel is located in are considered, then among these only the maximum water depth, its RP and its associated river pixel are saved.

This modification was introduced for two main reasons. First, rivers frequently form administrative boundaries between regions or countries, making the selection of which level of protection to apply, if neighboring regions have different protection standards, ambiguous. Second, hazard tiles can extend far downstream, sometimes up to 200 km from the originating river pixel (Figure 10) and therefore excluding an entire tile based solely on the protection level of the region containing the river pixel would ignore the possibility that downstream regions with lower protection standards could still be affected by flooding originating from the same river pixel.

After processing all peaks associated with an event, the maximum flood extent and maximum water depth maps were obtained, together with information on the river pixels responsible for flooding each floodplain location and the corresponding return period.

Population exposure and economic damages were then calculated. The affected population was estimated as the sum of the population living in all pixels where water depth exceeded zero. Economic damages were estimated using the land-cover class of each inundated pixel and the corresponding NUTS3 specific depth–damage curve, following the methodology described in Dottori et al. (2017, 2023); Rojas et al. (2013). Results were aggregated at NUTS3 level but stored at both pixel and regional resolution. Damage estimates were also disaggregated by the five land-cover groups considered in this study. The same procedures were applied to both simulated flood events and satellite-derived (observed) GFM flood events.

Clipping on basins

To enable consistent comparisons between simulated and observed flood events, a common spatial unit was required. The clustering procedure used for the generation of the GFM events in fact followed a different workflow (Betterle and Salamon, 2025), making direct comparison event by event challenging.

Therefore, both simulated and observed flood events were spatially partitioned using the level-2 hydrological basins defined by FAO (2021). Each event was clipped to the boundaries of these basins, used as the common spatial unit for subsequent comparisons between simulated and observed floods.

2.5 Matching of Events, Misses and False Alarms

Matching of Events

Two types of event matching were performed in this study. The first between simulated flood events and reported events from the Hanze database, while the second between simulated events and observed events derived from the GFM dataset.

The matching procedure with the Hanze dataset was conducted between all reported events and the subset of simulated events with total damages exceeding the minimum damage reported in Hanze (i.e., simulated damages > 0.1 million €). In contrast,

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

the comparison with GFM observations was performed using simulated and observed events clipped to hydrological basins, after filtering out events with flooded areas smaller than 1 km^2 in both datasets.

This threshold was mainly introduced to exclude very small clusters of flooded pixels in the GFM dataset that are unlikely to represent actual river or flash floods. In fact, despite such events accounting for approximately 51 % of all GFM events (and between 20 % and 35 % of simulated events depending on the protection scenario), they contribute only about 4 % of the total damages in the case of GFM and less than 1% for simulations. Therefore, despite the high number of events with areas smaller than 1 km^2 in the GFM, their impacts are negligible, which is even more so for the simulations. Many of the small clusters in the GFM can in fact likely be associated to water accumulation related to irrigation (Betterle and Salamon, 2025), localized precipitation events or classification errors related to similar radar back-scattering signatures (Risling et al., 2024). For more information and considerations regarding the effect of the thresholds, refer to Sections A6 and B5.

The matching between simulated and Hanze events is based on their temporal and spatial overlap. Each simulated event is compared with all reported events in the Hanze dataset and two events are then considered a match if:

1. They affect at least one common NUTS2 region or level-2 hydrological basin, and
2. Their start and end dates overlap by at least one day, allowing for a tolerance window of ± 1 week around the actual start and end dates.

A single reported event can potentially match multiple simulated events and vice versa. For this reason, matched events are later grouped into clusters. The matching and clustering operations were implemented using a graph-based approach allowed by the Python $networkX$ package.

The matching procedure between simulated and observed GFM events is instead more restrictive. In addition to the temporal and spatial criteria described above, a third condition must also be satisfied. Specifically, at least one river pixel responsible for the simulated flood must also be capable of flooding at least one pixel of the observed flood extent. This relationship is determined using the flood-hazard tiles associated with the 500-year return period, which represent the maximum potential influence area for each river pixel. If all three conditions are satisfied, the simulated and observed events are considered matched, then, as in the previous case, clusters of matched events are subsequently grouped together.

One limitation of this approach is that both simulated and observed events were previously clipped to hydrological basins. As a consequence, the spatial extent of matched events may be smaller than that of the original simulated events. Similarly, observed events originating from the same flood tile may remain unmatched if they affect a basin where no simulated flood occurred. However, this basin-based comparison also improves the comparability of simulated and observed events in terms of spatial extent and associated impacts. Section A6 reports the matching algorithms in greater detail.

False Alarms

After removing matched events, simulated events that do not correspond to any GFM event are classified as false alarms. Two categories of false alarms are defined.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

1. If the simulated event satisfies the temporal and spatial conditions described above (same basin or region and similar time window) but does not share any common river pixel capable of flooding the observed event, it is classified as a near-match false alarm, or
2. If no observed event occurs within the same basin or region and time period as the simulated event, it is classified simply as a false alarm.

Misses

Missed events (i.e., observed events without a corresponding simulated or reported event) are identified using a slightly different procedure. Instead of comparing observed events directly with simulated flood extents, the comparison is performed with the underlying river discharge time series, in terms of return periods. For each unmatched observed event, all river pixels capable of flooding any observed pixel are first identified using the flood hazard tiles. Then, within a tolerance window of ± 1 week around the observed event period, the discharge return periods at these river pixels are examined: if the maximum discharge return period at any of these river pixels exceeds the flood protection level of any observed floodplain pixel, the event is not classified as a miss. Otherwise, misses are further classified according to the underlying cause:

- Protection-related misses: if a river pixel could potentially flood the observed area with a return period higher than 10 years but lower than the corresponding protection level.
- Hazard-map limitation misses: if the maximum return period of any of the river pixels that could potentially flood the observed area lies between 2.5 and 10 years, indicating that flooding could occur, but falls below the minimum return period for which flood hazard maps are available.
- General misses: if none of the relevant river pixels exceed a return period of 2.5 years during the event window.

Finally misses, false alarms and matches were also stratified according to catchment size, defined as the median upstream drainage area of all river pixels responsible for the simulated flood (in the case of simulated events) or of all river pixels capable of flooding the observed area (for GFM events).

Catchment size classes were defined to distribute simulated and observed events relatively evenly across classes while still representing meaningful differences in basin size.

2.6 Evaluation of Results

Beyond event specific comparisons, results were also evaluated using several metrics, which are reported and explained here.

Flooded Area Metrics

Matched simulated and observed events were compared in terms of total damages, total flooded area, and spatial overlap between simulated and observed flood extents. Spatial agreement was evaluated using two widely used metrics: the True Positive Rate (TPR, or Hit Rate, Equation 2) and the Positive Predictive Value (PPV, Equation 3).

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table 1. Descriptions of the contingency table elements.

	Wet in GFM	Dry in GFM
Wet in Simulations	True Positive	False Positive
Dry in Simulations	False Negative	

The TPR measures the tendency of simulations to underestimate flooded areas, whereas the PPV measures the tendency to overestimate them. In an ideal case of perfect agreement, both metrics would be equal to 1 (Wing et al., 2017). Table 1 summarizes the three possible pixel classes used in these calculations. The true negative class is not included because there is no unique definition of a true negative pixel in this context: identifying such pixels would in fact require specifying an external bounding region that represents the maximum possible flood extent for each event.

$$TPR = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

$$PPV = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

Risk Estimate Agreement

To evaluate the overall consistency of flood risk estimates across different datasets and protection scenarios, we used Expected Annual Damages (EAD) over the period 2015–2024 as a proxy for flood risk in each NUTS region. To enable meaningful comparisons across datasets, the EAD values were categorized into five classes corresponding to the quintiles of each dataset's distribution. These five risk categories are defined as follows: *Q1: Very Low Risk, Q2: Low Risk, Q3: Medium Risk, Q4: High Risk, Q5: Very High Risk*. Cross-dataset agreement was assessed by comparing the assigned risk category of each NUTS region across datasets. For this purpose, we employed the Cohen quadratic weighted Kappa statistic, an extension of the widely used Cohen Kappa (Tehrany et al., 2015; Vergni et al., 2021; Trigg et al., 2016). This metric quantifies the similarity of categorical assignments while accounting for chance agreement (Ben-David, 2008). Unlike the standard Kappa, the quadratic weighting allows the analysis to appropriately reflect ordinal differences between categories, so that misclassifications between adjacent risk classes are penalized less than misclassifications between categories at opposite extremes.

The standard Cohen Kappa (Cohen, 1960) expresses the level of agreement between two classified variables, and it is defined as

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{4}$$

where p_0 is the observed agreement ratio and p_e is the expected agreement if classes were randomly assigned. In the case of the weighted Kappa with quadratic weights, the integer distance between classes (e.g. the distance between Q1 and Q2 would be 1, while that between Q1 and Q4 would be 3) is also taken into consideration through a weights matrix, which is multiplied

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

by the expected and observed agreement. Weights are calculated as follows:

$$w_{i,j} = 1 - (\frac{i-j}{n-1})^2 \tag{5}$$

with i and j being the (increasing) numerical category values and n being the number of different categories (5 in this case) (Ben-David, 2008). The Kappa statistic ranges from 0 (no agreement beyond chance) to 1 (perfect agreement). Negative values indicate instead agreement worse than those expected by chance. Using this approach, we assessed whether, despite differences in estimated damages and flooded areas, the relative ranking of NUTS regions was consistent across datasets or, in other words, whether regions identified as the most or least at risk were consistently recognized across datasets.

The analysis included simulated flood events under different protection scenarios, GFM-based observations, and reported impacts from the Hanze dataset. In addition, risk estimates derived from the static composite flood hazard maps (Section 2.1) were included for comparison. For these composite maps, EAD was calculated by converting water depths to damages using the respective depth–damage curves and then weighting the damages by the probability associated with each return period, as described in Equation 1. This choice allowed for additional comparisons between both simulated and observed floods with the average expectation deriving from the same original data of the simulations, which turn out especially useful to assess whether the impacts in the last 10 years have been, for each region, higher or lower than expected.

3 Results

3.1 Historic Impact Assessment

The choice of a flood protection dataset strongly influences the simulated impacts in several ways: it changes the total number of floods that get simulated, altering in turn the total impacts, which over long periods of time might diverge significantly. While entirely avoiding the introduction of flood protections, and in part also considering the P25 protections, causes a large number of simulated floods especially with lower, more frequently exceeded RP, employing FLOPROS protection changes not only the total number of floods, but also their distribution across return periods (Figure 3). Not accounting for protection or accounting for limited protection levels (P25) in fact leads, for the majority of countries, to shares of floods with low return period (return periods of 10 to 30 years) between 50% and 80%, while with FLOPROS these shares would decrease to between 20% and 40%. These differences are aligned with the actual protections levels and differences are more pronounced where protections are very high in the FLOPROS dataset, like in Poland, Germany, the Czech Republic, Slovakia and Hungary, and for larger countries like France and Spain. For more information regarding the location of floods refer to Section B2.

For most countries (except Montenegro, Switzerland, Ireland and Luxembourg), even under the No Protection setting, the share of floods with return periods larger than 100 years is between 10% and 15%, a quite concerning finding given the relatively limited length of the study period of 33 years and the catastrophic consequences of floods of this kind, even if on minor rivers.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

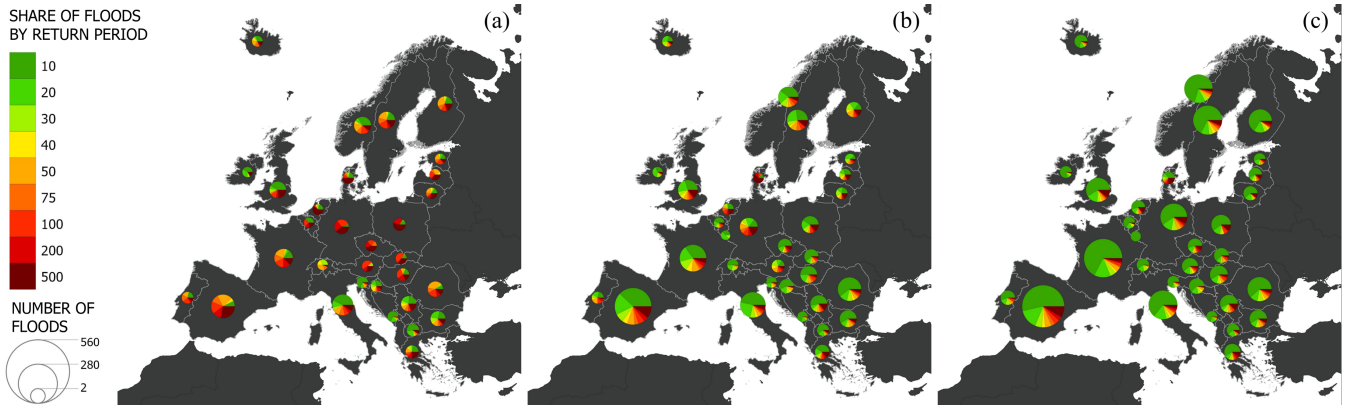

Figure 3. Number of floods and their distribution among return periods divided by country for the FLOPROS (a) and P25 (b) datasets and with No Protections (c).

Regardless of the protection dataset, damage hot-spots can be identified along the Alps and the Po' river basin, in the entirety of the Czech Republic, along the lower course of the Danube, in the border regions between Poland and Germany, and in central France (Figure 4).

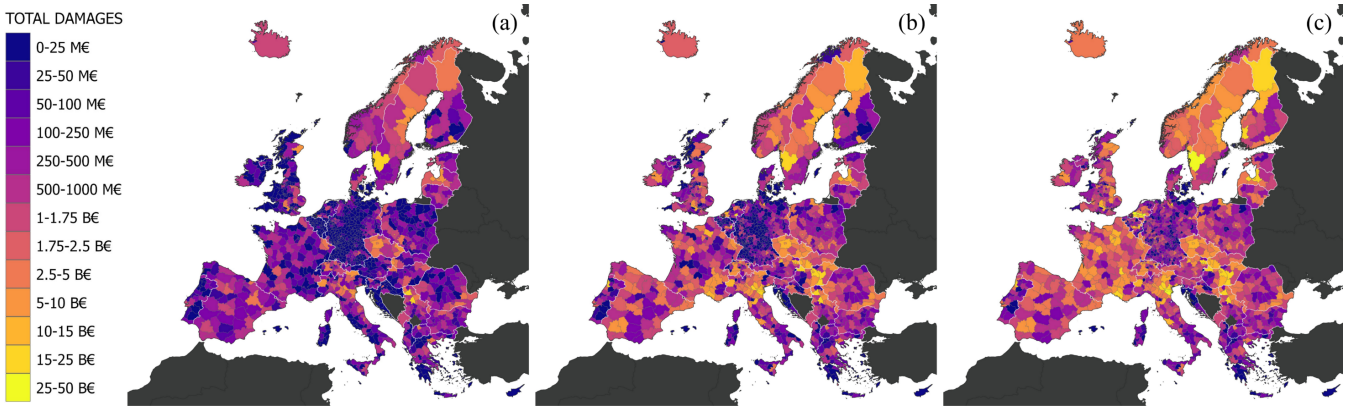

Figure 4. Total simulated flood damages from 1992 to 2024 with FLOPROS protections (a), P25 protections (b), and No Protections (c).

The difference in total damages (in the period 1992-2024) obtained with different protections is significant, especially for the whole of eastern Europe, where damages obtained with FLOPROS are between 2 and 10 B €lower than those deriving from the use of P25. Moving to the case without protections instead causes a further increase in damages, especially in France, the UK, and Northern Italy (Figure 4, for more information see also Figure B6). A peculiar case is Scandinavia and Lapland where results obtained through the use of FLOPROS are more than 10 B€lower than those obtained with P25 and more than 15 B€lower compared to the No Protections case, despite the difference in protection not being dissimilar to that of many other regions.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Comparing the simulated total damages over the entire study period (1992-2024) with reported impacts from different sources - the Disaster Risk Management Knowledge Centre (DRMKC), Hanze and the European Environmental Agency (EEA) - reveals a consistent overestimation of damages across all protection settings. Despite representing the dataset with the highest protection levels, FLOPROS still yields total damage estimates approximately twice as large as those reported by Hanze and the DRMKC, and about 170 % of the values reported by the EEA (Table 2). Furthermore, even under the highest protection setting, the number of simulated flood events remains about 30 % higher than the total number of reported events, indicating a systematic overestimation of both flood occurrence and associated impacts.

Table 2. Comparison of flood damages and number of events across flood protection and impact datasets (damages exclude the UK, Serbia, Montenegro, Albania, and Switzerland).

	FLOPROS	P25	No Protections	DRMKC (Floods)	Hanze	EEA (Floods)
Total Flood Damages (B€) (1992–2024)	568.4	1712.2	2695.8	209.6	250.6	334.9
Total Number of Events (1992-2024)	1680	2958	4137	1197	1251	

Country-level analysis shows that the choice of the protection dataset affects not only absolute damage estimates but also the ranking of the most affected countries. For example, using P25 instead of FLOPROS increases damages by more than 500% for France and Hungary, and by over 1200% for Austria and the Netherlands. Despite these variations, Italy, the UK, France, Germany, and Poland consistently rank as the most impacted nations in terms of total damages (Figure 5). Median estimates indicate that P25 increases damages by a factor of 2.6 relative to FLOPROS, while neglecting flood protections entirely results, in the median, in a further 1.3 fold increase. These results highlight the critical influence of protection assumptions in large-scale flood risk assessments, not only in terms of absolute damage estimates but also with respect to their relative distribution across countries, as will be further discussed in Section 3.4.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

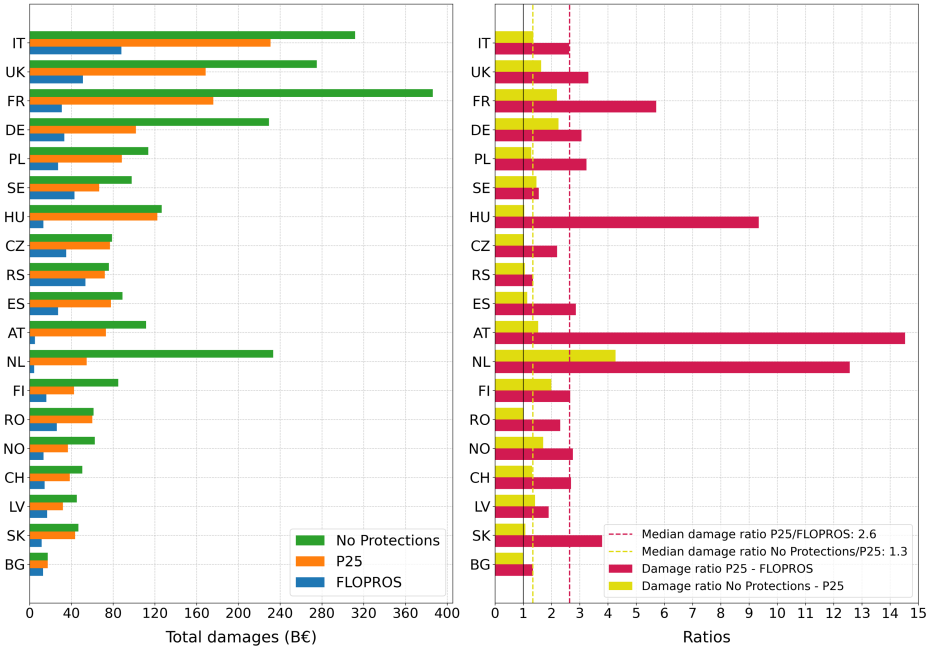

Figure 5. Total damages (a) and damage ratio (b) by country obtained using FLOPROS, P25 and No Protections for the 19 most impacted countries, ranked by average rank. For more information see Section B2.

Matching Simulated and Reported Events

Beyond total impacts, we evaluated to which extent simulations reproduce the same set of flood events reported in the Hanze dataset (with the methodology explained in Section 2.5).

Results again vary depending on the protection dataset that is used, but two consistent trends emerge: an increase in the share of matched Hanze events and a decrease in the share of matched simulated events with a decrease in protection return periods. Even in the No Protections setting, only 45% of Hanze events can be matched, corresponding to 53.7% of events with reported impact data, indicating moderate overall agreement between the two products (Table 3). Using P25 protections instead of FLOPROS increases the share of matches of Hanze by 13%, while reducing the share of matched simulated events by only 1.6%; not using any protection dataset instead only provides marginal benefits, increasing the share of matches by 3.2% but decreasing that of simulations by 4.3%. Overall, then the P25 dataset seems to be the best compromise when using Hanze as a reference. However, this can also be caused by the fact that P25 protection is very closely connected with Hanze, which has been used for its development (Paprotny et al., 2025)).

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table 3. Percentage of historic (1992-2024) events that were matched between simulations and Hanze with different protection datasets and without protections.

	FLOPROS	P25	No Protections
Simulations	24.9%	23.3%	19%
Hanze (Total)	28.8%	41.9%	45.1%
Hanze (With impact data)	36.6%	50.8%	53.7%

3.2 Comparison of observed, simulated and reported events

In addition to reported impacts, simulations were also compared to observed flood extents from a recently published dataset (Betterle and Salamon, 2025) derived from the GFM system of the Copernicus Programme (Matgen et al., 2020). As this dataset covers the period 2015–2024, simulated and reported events were restricted to the same time range. Matches, misses, and false alarms between simulated, observed and reported events were then identified according to the procedures explained in Section 2.5. Misses (i.e., events present in the observed event catalogue but not in the simulated nor reported catalogues) and false alarms (i.e., events present in the simulated event catalogue but not in the observed nor reported datasets) are always to be interpreted relative to the other datasets. In fact, we intentionally avoid defining a single ground truth with respect to the number, extent, or characteristics of floods and their impacts; hence, the following analysis should be considered as an assessment of differences between datasets rather than as a validation of either one against the others.

Matches, misses and false alarms

Concerning simulations, regardless of the protection setting, false alarms always account for 50-55% of all events, while misses are instead between 77 and 84% of observed events, the majority of which (58% of all events) being caused by the lack of river discharges with significant return periods (Table 4). This points either towards a fundamental misrepresentation of the return period of discharges or, more realistically, to a large presence in the dataset, of events that should not actually be considered floods but could instead represent water pools generated by other causes. Misses caused by the lack of flood hazard maps for RP lower than 10 years are also significant, accounting for around 20% of all observed events, while misses caused by too high protections are 5.7% of events in the FLOPROS case, which halve to 2.5% in the P25 case. Moving to matched events, for the simulations half of all matches are with both GFM and Hanze, while the remaining half is equally split between matches only with GFM and only with Hanze. Similarly, in the GFM case around half of all matches are with both other datasets, except in the case of FLOPROS protections, where only 4.7% of observed events find a match with both simulations and Hanze, a share that doubles if P25 protections are used. At the same time moving from FLOPROS to P25 and to the No Protection case, increasingly reduces the number of GFM events matched only with Hanze (by 2.8% and a further 1.3% respectively) and increases the share of events matched with both simulations and Hanze.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table 4. Distribution of matches, false alarms, and misses by type and under different protection datasets. Percentages refer to the share of simulated and observed events clipped to basins that belong to each miss/false alarm or matches class. Every match, false alarm or miss type is exclusive (i.e. simulated events matched only with GFM are in the "Matches GFM" class, while those matched with GFM and Hanze are only in the "Matches GFM & Hanze" class.

	Simulations						GFM						
Protection dataset	Matches GFM	Matches Hanze	Matches GFM & Hanze	False alarms	False alarms Near match	Number of events	Matches Sim	Matches Hanze	Matches Sim & Hanze	Misses RP 2.5–10	Misses RP < 2.5	Misses Protection error	Number of events
FLOPROS	12%	11.6%	22%	54%	0.4%	986	1.7%	9.5%	4.7%	20.6%	57.9%	5.7%	
P25	11.8%	12%	25.3%	50%	0.7%	2087	3.7%	6.7%	9.4%	19.6%	58.2%	2.5%	8684
No Protections	13.7%	10.4%	24.7%	50.6%	0.5%	2863	5.6%	5.4%	12.2%	18.6%	58.3%	0%	

In terms of damages and flooded areas, both misses and false alarms represent a significant share of the total. In the GFM case, misses account for 68% to 79% of the total flooded areas and for 53% to 65% of the total damages, with the majority of these being misses caused by the lack of significant discharges, while in the case of simulations false alarms represent around 30% of both flooded areas and damages (Figure 6). In addition there is a much larger difference between flooded areas and respective damages from the GFM compared to the same difference in the simulations (the ratio between flooded areas and damages is around 20 times higher for the GFM than for the simulations), indicating that while total flooded areas are relatively comparable between the two data sources, total damages are not, with damages estimated from the observations consistently being much lower than those derived from the simulations, regardless of the protection setting. In fact while total flooded areas from the GFM are between 0.9 and 2.8 times those of the simulations, in the case of flood damages GFM estimates account for only 0.04-0.17 times the total damages of the simulations.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

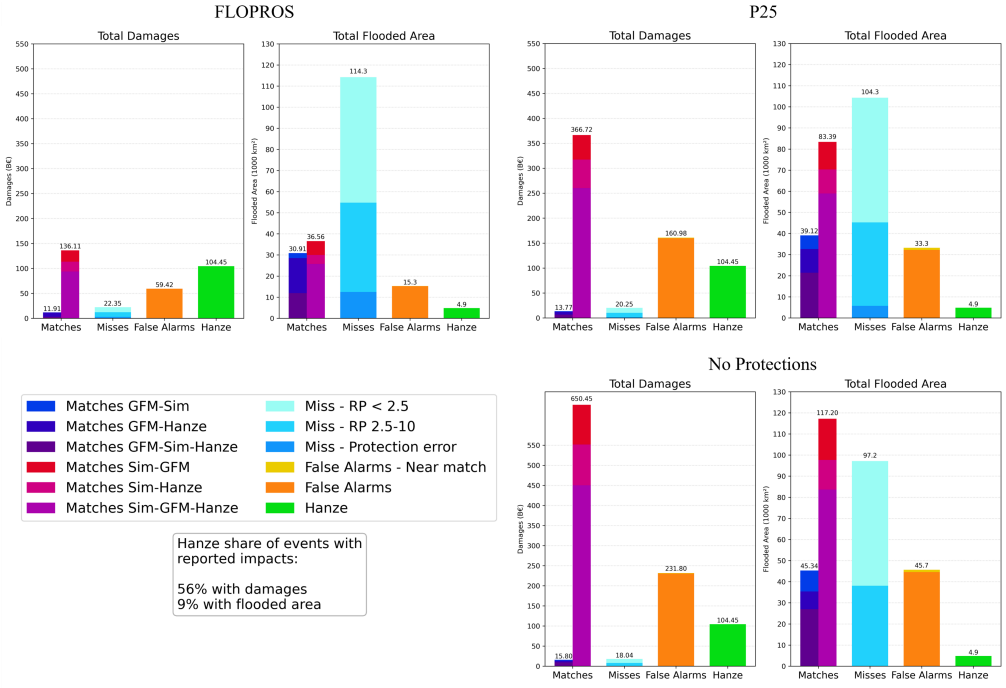

Figure 6. Total damages and flooded areas for the two protection datasets, divided by match, miss, false alarm, and their different types.

Dividing matches, misses and false alarms by catchment size reveals that the share of matches from the simulations increases with catchment area, particularly that of matches with both other datasets, while the share of matches between GFM and Hanze decreases, as well as the proportion of misses caused by excessively high protection levels (Figure 7). The latter trend can be easily explained by the fact that larger rivers typically have higher protection standards than smaller rivers within the same administrative regions. The former trends instead suggests a greater capability of the EFAS-LISFLOOD modelling framework to reproduce floods in larger catchments, especially those exceeding 45,000 km^2, compared with smaller basins.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

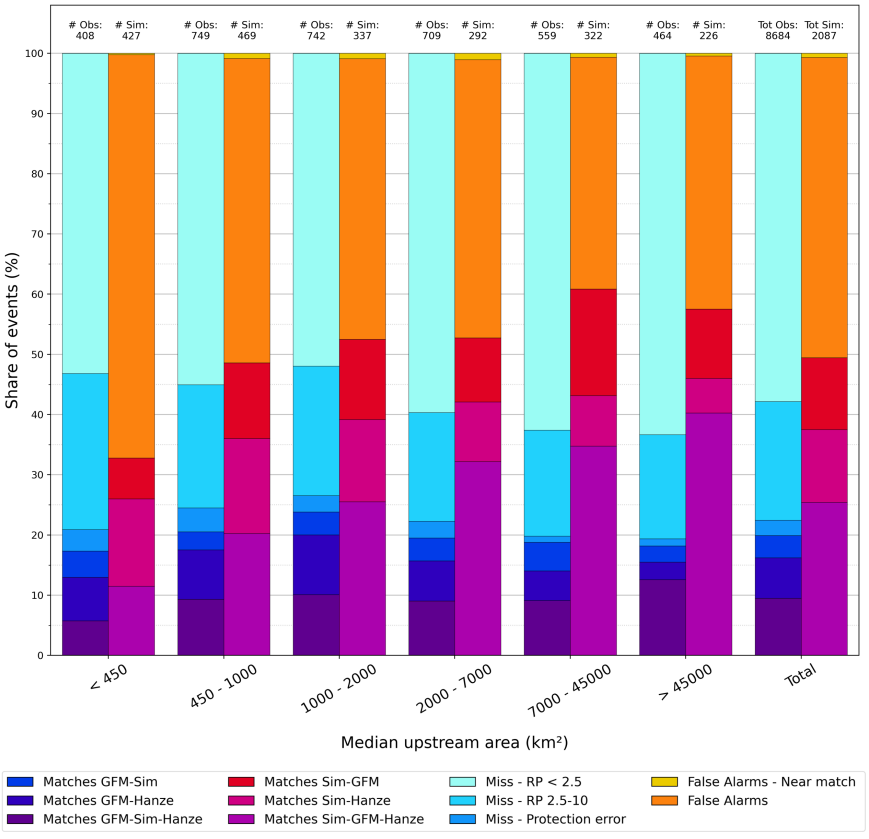

Figure 7. Share of matches, misses and false alarms by their respective type between observed floods and P25 protection estimates, divided by median upstream area class. Results for FLOPROS and No protections are reported in Figure B9

Matched flood events

Overall 139, 250 and 322 separate clusters of flood events are found in common between GFM and simulations with FLOPROS, P25 and No Protections respectively, of which only between 37% and 52% are present also in Hanze. Considering common floods between Hanze and simulations instead, around 40% of them are discarded when accounting for GFM as well (Table 5). In terms of total number of common events among all three datasets that occurred between 2015 and 2024, 55 separate floods are found when considering FLOPROS protections while 82 and 83 floods are found with the P25 and No Protection settings respectively.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table 5. Total number of matched events (after clustering) for each combination of datasets and their share when considering a third dataset, for each protection setting. The number of matches between all datasets (Sim-GFM-Hanze) is smaller because the introduction of a third dataset may connect two previously distinct events.

	FLOPROS			P25			No Protections		
	# Events	of which in Sim-GFM-Hanze	# Events Sim-GFM-Hanze	# Events	of which in Sim-GFM-Hanze	# Events Sim-GFM-Hanze	# Events	of which in Sim-GFM-Hanze	# Events Sim-GFM-Hanze
Sim–GFM	139	72 (51.8%)		250	117 (46.8%)		322	121 (37.58%)	
Sim–Hanze	101	59 (58.4%)	55	145	91 (62.76%)	82	152	94 (61.84%)	83
GFM–Hanze	132	60 (45.5%)		132	84 (63.64%)		132	87 (65.91%)	

Some possible sources for the inconsistencies between datasets can be hypothesized by analyzing the location and extent of matched events between any two datasets that are not present in the third dataset. In the case of matched events between GFM and Hanze not present in the simulations, they tend to be clustered in France and England, with minor hot-spots in Romania, Northern Italy, Southern Spain, the Baltics and Ireland. When protection datasets are not considered, the number of unmatched events in France, England and Romania decreases dramatically, indicating that in those regions and for the rivers in question, flood protections are very likely overestimated, especially in the FLOPROS dataset. In the other regions discrepancies can instead be connected to the underestimation of discharge RP and the lack of flood hazard maps with RP lower than 10 years (Figure 8). Events present both in simulations and Hanze but not in GFM are mainly located in Southern Europe, especially Italy and Spain, but with some noticeable clusters also in Sweden, Poland, Serbia and France. In these cases it is likely that the flood dynamics were very fast and difficult to capture by satellite, an hypothesis supported by the fact that in this case mainly minor rivers are identified. Lastly, events present in both GFM and simulations but not in Hanze are more spread across the continent, but with a higher presence in Scandinavia and Eastern Europe, together with Spain. However when lower protections are considered or when they are not accounted for, also France, the UK and Ireland start to exhibit several unmatched events with Hanze. The large number and extent of events present in both GFM and simulations but not Hanze in Scandinavia could be explained by the sparseness of population centers and human activities in general in those areas. In fact Hanze mainly relies on governmental reports, news, and scientific papers in order to identify floods (Paprotny et al., 2024) and these are much less likely to be produced in areas where impacts on human activities could have been very low or null, even if floods did actually occur. However this is unlikely the case for France, the UK, Italy, Spain, Hungary and Romania, where causes for the discrepancies should be found elsewhere.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

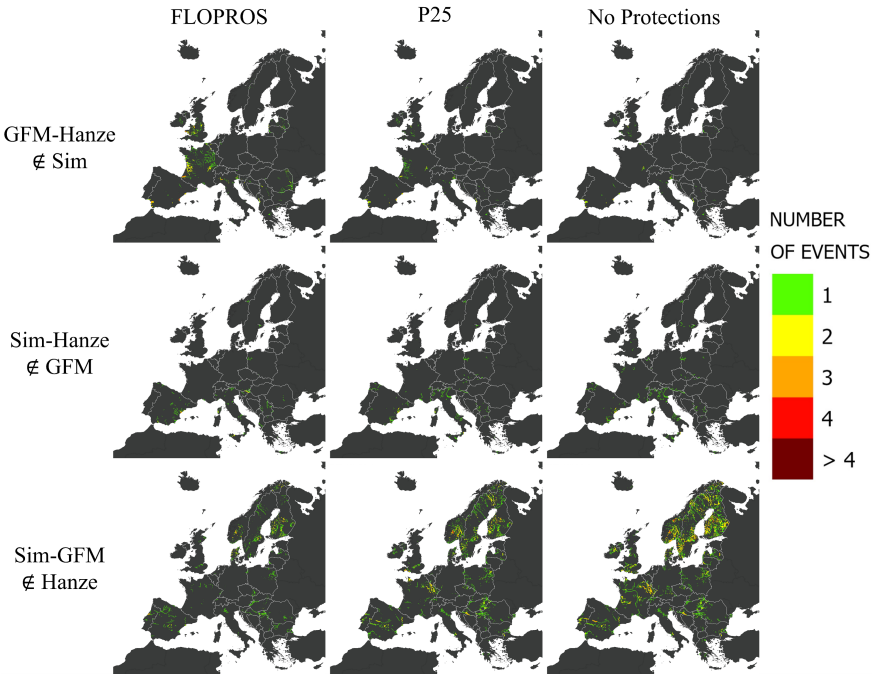

Figure 8. Location, extent and number of matched events between any two dataset combination not present in the third dataset, with the different protection settings. The extent and number of events in the Sim-GFM case account for both simulated and observed flood extents. The location and extent of matched events between any combination is instead reported in Figure B10.

These results highlight several important points: the first is that while for both simulations and GFM the highest share of matches is found when matching on both other datasets, these events still represent only a quarter of all events in the case of simulations, and less than a tenth in the case of GFM. In addition, even when considering matches with just a single other data source, 50% of simulated events (accounting for 30% of damages), 60% of observed events (accounting for more than 53% of damages) and around 50% of reported events still do not find a correspondence with any other dataset, highlighting the large inconsistencies between all three datasets. The second important result that we found is that the use of FLOPROS as protections dataset severely limits the number of GFM events that find a match with simulations, even if they are already present in both GFM and Hanze, and while this is true also for the P25 dataset, the issue is less evident. In addition P25 also allows to obtain the lowest share of false alarms of all protection settings. In terms of separate common flood events, the use of the P25 protections allows to identify 27 more common flood events between all three datasets compared to the case with FLOPROS protections, an increase by 50%. However if flood protections are not accounted for, only a single additional flood event is found, despite a significant increase in the absolute number of matched simulated and observed events. In other words not accounting for protections has the main effect of increasing the overall size of the flood, both in terms of simulated area and of observed area matched with the simulations, but without having a significant effect on the number of separate floods that can be identified. Despite certainly not being a perfect representation of flood protections then, the P25 dataset might

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

still represent the best compromise when assessing flood impacts through EFAS simulations. A third important point worth highlighting is that the availability of flood hazard maps for return periods between 2.5 and 10 years would allow to more than double the number observed events that can find a match with simulations, protections allowing, and for this reason we recommend making these available in the next versions of EFAS.

Overall these results show the importance of considering multiple data sources when assessing flood impacts, given the limitations of each, as well as the significant consequences of the choice of a flood protection dataset instead of another and highlight the fundamental differences between damage estimates from GFM observations and from EFAS simulations.

3.3 Analysis of Matched Events

To better understand the sources of the discrepancies between simulations and GFM, matched events and their characteristics were analyzed further. In this section we consider all events matched between observations and simulations, due to the larger sample sizes, while results of events matched also with Hanze are reported in Section B4 of the Supplementary Information. Matched events were first grouped into classes based on observed and simulated flooded area size and then flooded area bias and overlap area were assessed between simulations and observations.

Simulated flooded areas bias

The multiplicative bias between simulated and GFM flooded extents was evaluated for each class. This analysis reveals two contrasting trends: an increase in bias with increasing simulated flooded area, and a decrease with increasing observed flooded area from GFM (Figure 9). Correlation analysis confirms that both relationships are statistically significant, although the positive correlation between bias and simulated flooded area is stronger (0.6, Table 3.3). For events with small simulated flooded areas, the bias is often lower than 1, indicating that simulated floods tend to underestimate the extent of smaller events compared to observations. Conversely, the bias increases progressively for larger simulated floods, suggesting a tendency to overestimate the extent of large events.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

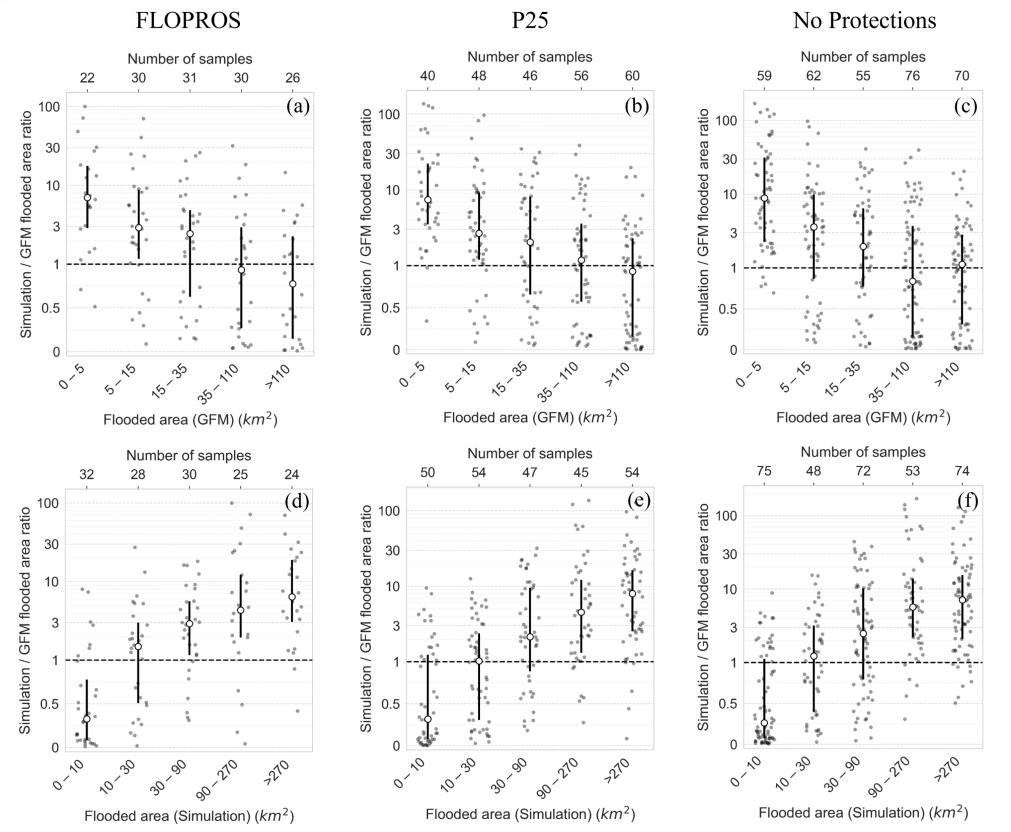

Figure 9. Bias of simulated flooded areas with respect to observed flooded area for FLOPROS (a,d), P25 (b,e) and No protections (c,f), divided by classes of observed flooded area (a,b,c) and simulated flooded area (d,e,f). Bias of estimated damages are reported in Figure B11.

At the same time, the decreasing bias observed with increasing GFM flooded extents—reaching a median value close to 1 for observed flooded areas larger than 35 km^2—suggests that the GFM-derived flood extents themselves may be underestimated, particularly for smaller events.

Table 6. Spearman correlation and 5-95% confidence intervals between bias in simulated flooded area and observed flooded area by simulated / observed flooded area size.

Total flooded area	FLOPROS	P25	No Protections
Simulation	0.61 [0.51, 0.69]	0.60 [0.52, 0.66]	0.61 [0.54, 0.66]
GFM	-0.47 [-0.58, -0.36]	-0.45 [-0.53, -0.37]	-0.42 [-0.49, -0.34]

These combined trends indicate that discrepancies between the two datasets arise from both an underestimation of small floods in simulations, together with a potential under-detection of small flooded areas in satellite observations, and from an

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

overestimation of the size of large floods from the simulations, especially those on major rivers. One possible cause for the overestimation of large floods by the simulations can be found in the flood hazard tiles and more specifically in their size.

For large rivers in fact each location in the floodplain could potentially be flooded by several river pixels, up to around 100-150, which can also be located very far upstream along the river, by more than 100 km and even more than 200 km in some cases (e.g. Danube, Rhine, Vistula and Elbe). This is a situation found on all major European rivers, especially on reaches located in plains. This fact justifies the large extents of simulated floods originating from these rivers, even when observed extents wouldn't verify it, since the activation of just one river pixel could potentially already generate very large flood extents.

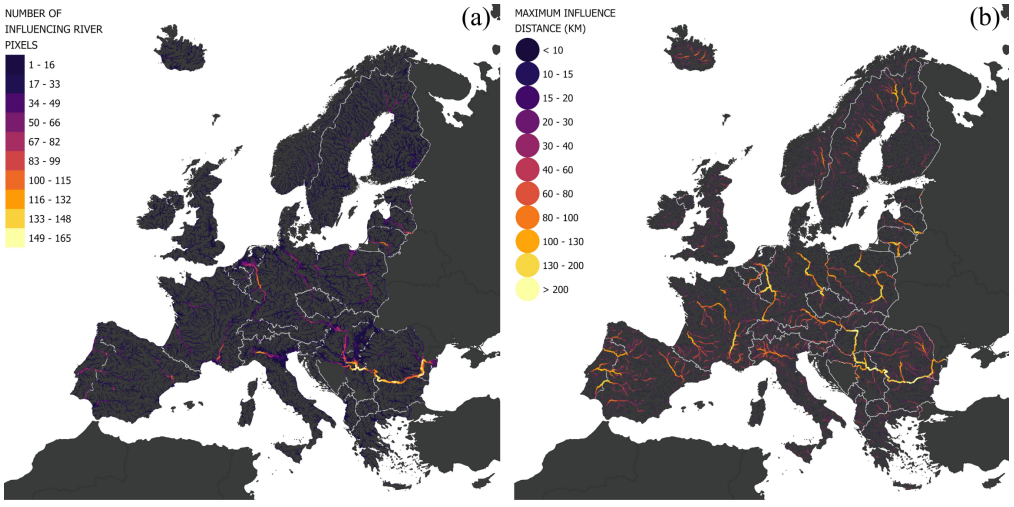

Figure 10. Number of river pixels influencing each floodplain pixel (a) and maximum influence distance of each river pixel (b), with a RP of 500 years

Overlap area between simulated and observed events

Analyzing the spatial overlap between simulated and observed flood extents highlights further inconsistencies. In terms of overlap metrics, only around 20% of events exceed a true positive rate (TPR) of 30%, while the proportion of events exceeding the same threshold for the positive predictive value (PPV) is even lower, at approximately 15%. TPR values are generally higher than PPV values, reflecting the overall tendency of simulated floods to produce larger flooded areas than the observed ones. In addition two trends are evident: TPR increases as the simulated flooded area increases, with a correlation of approximately 0.55, while PPV increases with increasing observed flooded area, with a correlation of around 0.45 (Table 7). Together, these patterns indicate that the relative overlap between simulated and observed flood extents improves for larger events. In other words, the agreement between simulations and observations tends to increase with flood magnitude. Nevertheless, overlap exceeding 30% of the observed flooded area occurs for only about 22% of events, and for only around 6% of events when measured relative to the simulated flooded area (for more information see Figures B12 and B13).

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

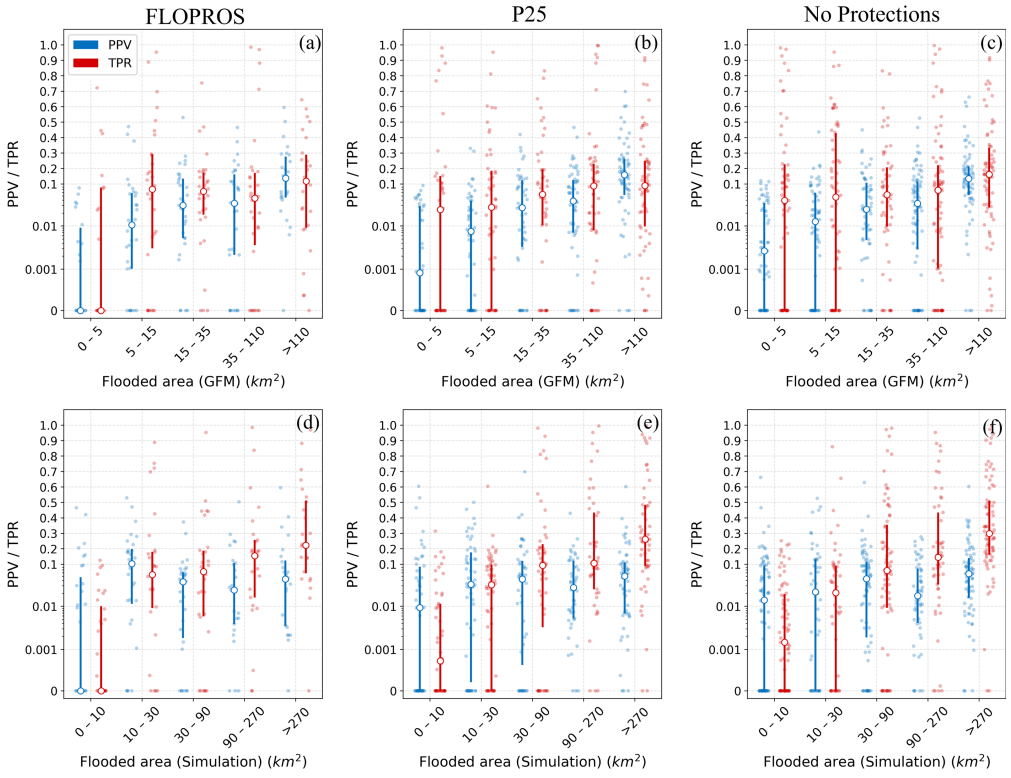

Figure 11. True Positive Rate (TPR) and Positive Predictive Value (PPV) of matched simulated and observed events with FLOPROS (a,d), P25 (b,e) and No protections (c,f) divided by observed area size (a,b,c) and flooded area size (d,e,f).

Table 7. Spearman correlation coefficients and 5–95% confidence intervals between PPV (and TPR) and simulated / observed flooded area.

	FLOPROS		P25		No Protections	
Total flooded area	PPV	TPR	PPV	TPR	PPV	TPR
Simulation	0.17 [0.02, 0.32]	0.53 [0.41, 0.62]	0.14 [0.03, 0.24]	0.55 [0.47, 0.62]	0.20 [0.11, 0.29]	0.58 [0.52, 0.64]
GFM	0.45 [0.32, 0.55]	0.16 [0.01, 0.30]	0.49 [0.41, 0.57]	0.17 [0.07, 0.28]	0.45 [0.37, 0.52]	0.14 [0.05, 0.24]

Water depth and affected land cover classes

The differences between flooded areas and estimated damages can be further explained by two main factors: differences in water depths between simulations and observations, and differences in the distribution of flooded areas across land cover classes. Simulated water depths are consistently higher than those derived from observations for all percentiles above the 5^{th}. Below this percentile, simulated depths are shallower because, while the observational dataset imposes a minimum water-depth threshold of 0.1 m, the simulated dataset allows any water depth value even below this threshold. The differences are especially pronounced in the central percentiles (25^{th}-75^{th}), where simulated water depths are substantially higher than observed values.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

For example, median water depths are 0.2 m in observations versus 0.65 m in simulations for the 25^{th}-50^{th} percentile range, and 0.38 m versus 1.5 m for the 50^{th}-75^{th} percentile range (Figure 12). These differences strongly influence the resulting damage estimates. Similar discrepancies occur at the highest percentiles, where simulated depths exceed 8 m while observed depths generally remain below 5 m. It should be noted, however, that satellite-based observations and the water depths derived from them, depend strongly on the timing of satellite acquisitions, which rarely coincide with the actual flood peak. In contrast, simulated flood depths and extents always represent peak conditions. Although both datasets describe the same phenomenon, neither one fully captures the dynamics of flood processes and the use of peak conditions in simulations versus snapshot observations is likely to significantly influence estimated impacts. A more comprehensive assessment of impacts would then require information not only on maximum water depth and flood duration, but also on the temporal evolution of water levels and, potentially, flow velocities.

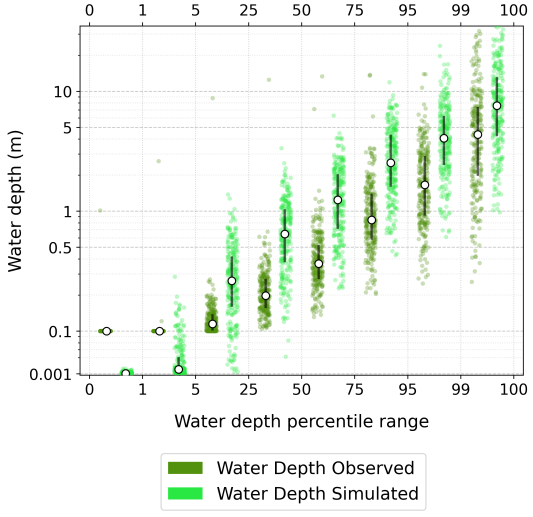

Figure 12. Simulated and observed water depth by percentile (P25 protections). Percentile ranges represent the edges of the data used to generate each plot.

Differences in land cover exposure also contributes to discrepancies in damage estimates. In both simulations and observations, agriculture represents the largest share of flooded area, followed by residential areas, industrial and business areas, and transport infrastructure (Figure 13). However, the simulations show a lower proportion of flooded agricultural land (74% compared to 83% in observations), with correspondingly higher shares in other land-use classes, particularly residential areas. Despite these seemingly small differences in flooded area distribution, the implications for damages are substantial, especially when combined with higher simulated water depths. For example, in the observational dataset, residential areas account for a median of only 0.26% of flooded pixels but contribute approximately to 15% of total damages. In the simulations, residential areas represent about 3% of flooded pixels, yet account for 38% of total damages. Similar patterns are observed for the industrial and

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

business sectors. Conversely, agricultural damages represent a relatively larger share in the observational dataset, where they account for 13% of total damages compared with only 1.2% in the simulations.

In addition the observational dataset presents a frequent absence of damage in the residential, industrial, and transport land-cover classes. Specifically, between 18% and 28% of events in the GFM dataset show no flooded areas, and consequently no damages, in these classes, whereas this occurs in at most 6–10% of the simulated events (for more information see also Figures B14, B15 and B16).

Overall, these results indicate that the observational dataset identifies a substantially larger share of flooded areas in agricultural land and a much smaller share in urban and infrastructure classes compared with the simulations. This pattern is consistent with known limitations of satellite-based flood detection methods Betterle and Salamon (2025). When combined with the lower observed water depths, the difference in exposure distribution helps explaining the inconsistency between relatively comparable flooded areas and much larger differences in estimated damages between simulations and observations.

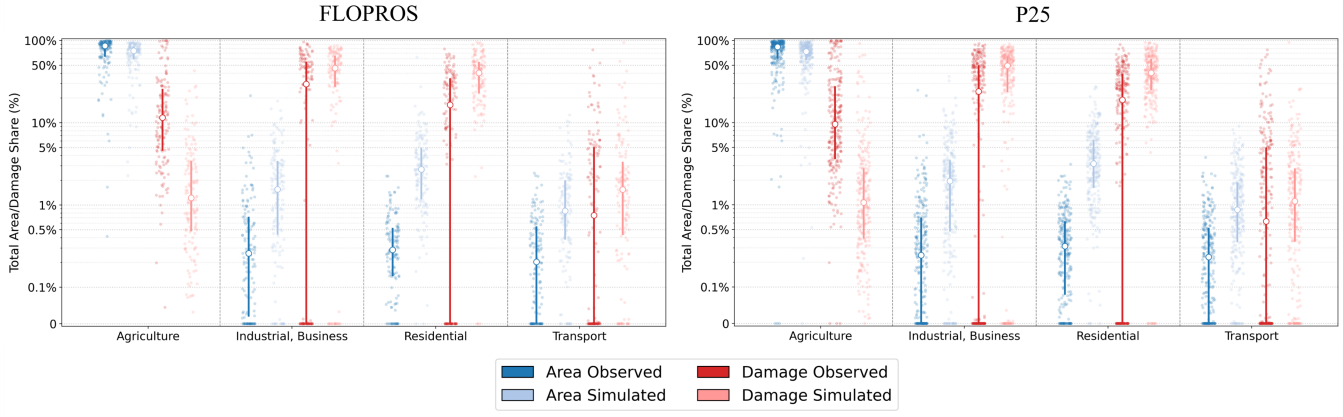

Figure 13. Shares of flooded area and estimated damages from matched event by land cover class and percentile range of water depth (P25).

3.4 Risk estimate agreement

As final step we assessed the overall similarity between the impacts and risk estimated through the various data sources and protection datasets. In this analysis we consider, in addition to simulations, observations and reported events, also EAD estimated through the use of (static) composite flood hazard maps.

The statistical agreement between risk categories across NUTS regions, assessed using the Cohen quadratic weighted kappa, highlights substantial variability at the regional level. This variability reflects the difficulty of consistently identifying highly impacted areas at finer spatial resolutions, particularly for the Hanze dataset, where impact data are aggregated at the event level (Figure 14). At both NUTS0 and NUTS3 scales, the highest levels of agreement occurs, as expected, between the datasets representing simulations with different protection levels and their corresponding static risk estimates derived from composite

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

flood hazard maps. Interestingly, using the higher FLOPROS protections reduces agreement with the corresponding static estimate more strongly than other protection settings when moving to the NUTS3 scale. This effect may be related to the relatively short analysis period (10 years), as higher protection levels reduce the number of simulated flood events, thereby increasing variability in regional risk estimates. A different pattern emerges when considering the observational dataset: in this case, the highest agreement is observed with the static risk estimate based on P25 protections, followed closely by the corresponding simulated estimate and the risk estimate calculated without accounting for flood protections. In contrast, the simulated estimate using FLOPROS protections shows a substantially lower agreement. This result suggests that, despite the potential overestimation of flood occurrences and impacts, the P25 protection dataset may represent flood risk more consistently than FLOPROS, even when used in combination with static flood hazard maps rather then simulations. Once again, the substantially lower kappa values obtained at the NUTS3 scale indicate that achieving agreement between observations and simulations becomes considerably more difficult at finer spatial resolutions.

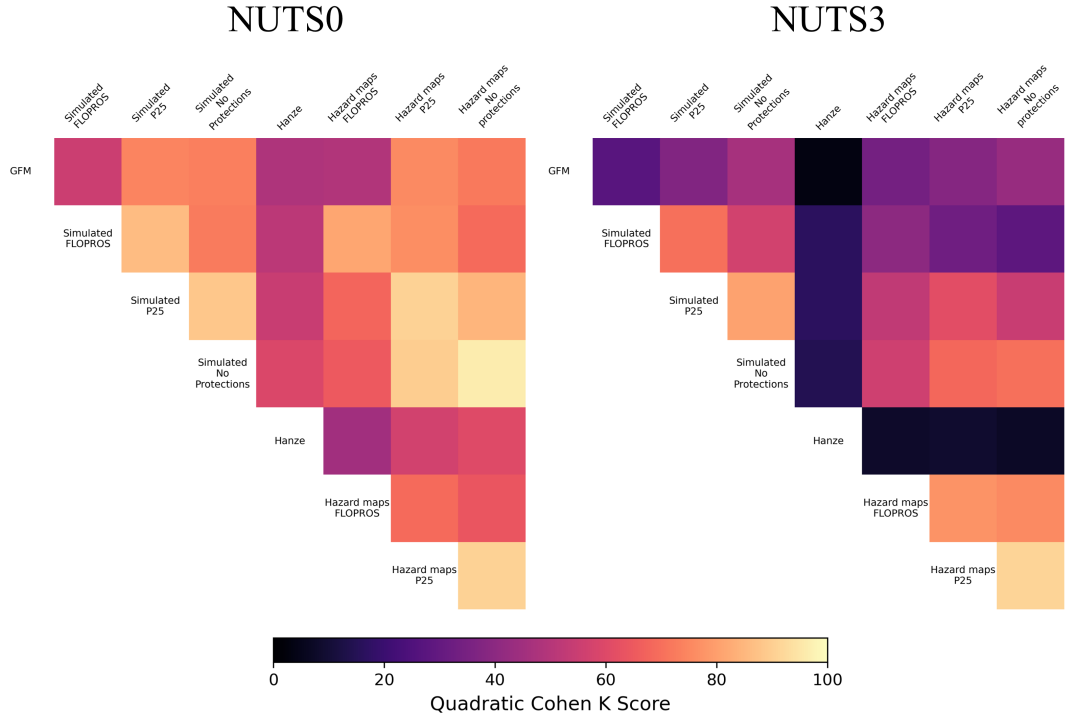

Figure 14. Quadratic Cohen K similarity score of EAD quintiles at NUTS0 and NUTS3 scale for all datasets.

For the Hanze dataset at the NUTS0 scale, the highest agreement is instead found with the static risk estimate that does not account for protection levels, which produces the most similar categorization of countries. At the NUTS3 scale, however, Hanze shows extremely low agreement with all other datasets, an expected result given the limited spatial resolution of the reported impact data.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Overall, the strong agreement observed between simulations and static risk estimates supports the validity of the analysis and suggests that composite hazard maps can provide meaningful risk estimates even for relatively short analysis periods. At the same time, very low agreement between simulations, observations, and reported impacts at the NUTS3 scale highlights that differences between datasets affect not only absolute impact estimates but also the qualitative characterization of flood risk and impacts across regions.

4 Conclusions

This study first provides a comprehensive assessment of how different flood protection datasets influence the results of large-scale flood impact modeling and then shows the degree by which simulated impacts compare with observed and reported flood events across Europe. The results highlight that the representation of flood protection is a primary driver of uncertainty, affecting not only the total magnitude of estimated damages but also the frequency, spatial distribution, and intensity of simulated flood events. Different protection datasets lead in fact to substantial variations in damage estimates, by up to an order of magnitude, and significantly alter the ranking of the most affected regions and countries. These findings demonstrate that flood protection assumptions are not a secondary modeling choice, but central to flood risk assessments performed via simulations.

At the same time, the comparison between simulations, satellite-based observations, and reported flood datasets reveals fundamental structural inconsistencies between data sources. Simulations tend to provide a physically consistent representation of flood processes but systematically overestimate both the number of events and their associated damages, particularly for large floods, at the same time failing in simulating several events that were both observed and reported. In contrast, satellite observations capture the spatial extent of flooding directly but are affected by limitations in detecting small or short-lived events, as well as by the timing of satellite acquisitions, which rarely coincide with peak flood conditions, and they are more prone to misidentification of flooded areas. Reported impacts are instead inherently incomplete and biased toward populated and economically relevant areas. As a result, large shares of events remain unmatched across datasets, leading to significant differences in overall impact estimates not only in quantitative, but also qualitative terms and at regional and even national scale.

The analysis also identifies several systematic sources of bias along the modeling chain. Simulations tend to overestimate the extent of large flood events, mainly due to the structure of flood hazard maps since floodplain pixels can be influenced by multiple upstream river pixels and over large distances. At the same time, events in smaller catchments more often represent false alarms, due to larger uncertainties in discharge return periods. Differences in estimated damages between simulations and observations are further amplified by discrepancies in water depth and land cover exposure: simulated water depths are generally deeper and affect a larger proportion of urban and high-impact areas, while observed floods are shallower and more concentrated in agricultural land. Combined, these effects explain why simulated damages are consistently much higher than those derived from observations, even when flooded areas are of a more comparable magnitude.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Among the tested protection datasets, the P25 dataset emerges as a pragmatic compromise. It provides a balance between reproducing reported and observed events, limiting false alarms, and achieving relatively coherent risk estimates across datasets. However, this improved performance may partly reflect its closer methodological linkage to reported flood data, rather than a more accurate representation of physical protection levels. Therefore, while P25 may be more suitable for applications requiring consistency with historical records, it should not be interpreted as a definitive representation of actual flood protection standards and, if used without the support of an observational dataset, can still allow for the simulation of a large number of events that likely never occurred.

Results further highlight several key data and modeling gaps. The absence of flood hazard maps for return periods between 2.5 and 10 years represents a major limitation, preventing the simulation of a large share of observed events. In addition, both modeling and observational approaches struggle to capture small-scale and fast floods, particularly in smaller catchments. Satellite-based observations are limited by detection thresholds, revisit times, and reduced sensitivity in urban or vegetated areas, while simulations are constrained by the use of static hazard maps representing peak conditions rather than full flood dynamics. Addressing these limitations will require both improved data availability and methodological advancements, including the development of dynamic modeling approaches employable at large-scales that better represent the temporal evolution of floods.

Based on these findings, several recommendations can be made for future work. First, the development and public availability of flood hazard maps for lower return periods should be prioritized, as this could potentially triple the number of observed events finding a corresponding simulated event. Second, greater integration between modeling and observational data is needed, leveraging their complementary strengths. Such integration could also make use of recent advances in model integration through AI. This integration could be particularly useful in urban areas and in small catchments, as well as for planning and insurance purposes. Finally, results underscore that currently no single dataset by itself should be considered a complete ground truth for flood impacts at continental scale and consequently future model validations should incorporate flood delineation data from multiple observational sources, such as different radar satellites or a combination of radar and optical sensors, in order to reduce ground-truth uncertainty and enable a more reliable validation of flood models.

From a policy perspective, these results demonstrate that flood risk estimates are highly sensitive to modeling assumptions and dataset choices, and should therefore be interpreted with caution. Rather than relying on single dataset estimates, historic risk and impact assessments should then systematically adopt multi-dataset approaches and explicitly account for uncertainty ranges incorporating information from various sources, as different modeling or dataset choices can lead to substantial biases in the identification of risk hot-spots and in the prioritization of risk mitigation measures.

In conclusion, this study shows that simulations, observations, and reported data provide complementary but inherently incomplete perspectives on flood risk. Differences are substantial and persistent, reflecting both methodological and conceptual limitations. Future efforts should therefore not only focus on identifying the best dataset or in improving modeling chains, but also on integrating multiple sources of information within a consistent framework that explicitly acknowledges and quantifies

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

uncertainty. Such an approach is essential for improving the reliability and credibility of large-scale flood risk assessments and allow more robust decision making.

Code and data availability.

The code, data and results of this work are publicly accessible through a Zenodo repository (Scarpellini et al., 2026). The repository also contains the simulation datasets with the three protection levels, which can be used freely in other research works.

Appendix A: Supplementary Methods and Data

A1 Flood Protection Datasets

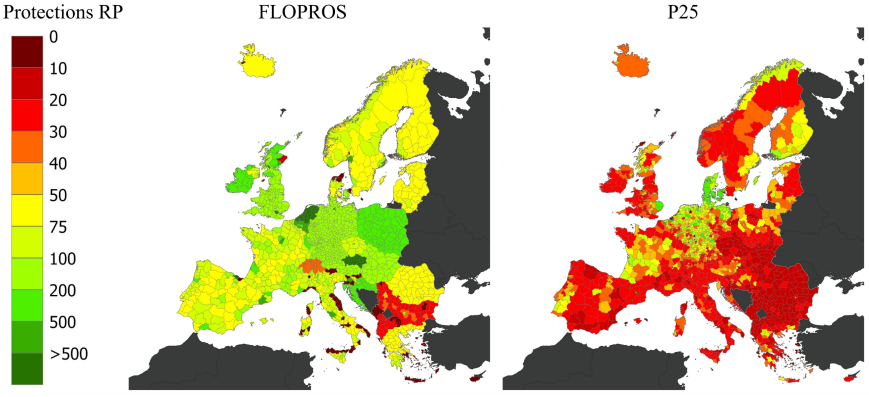

Figure A1. Protection in RP terms for FLOPROS and P25 datasets at NUTS3 resolution.

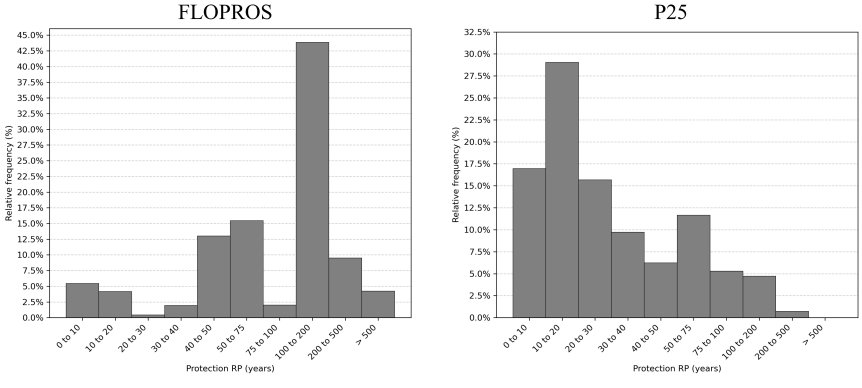

Figure A2. Distribution of FLOPROS (merged layer) and P25 (mean of the period 2015-2020) protections expressed in terms of RP and at NUTS3 resolution.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

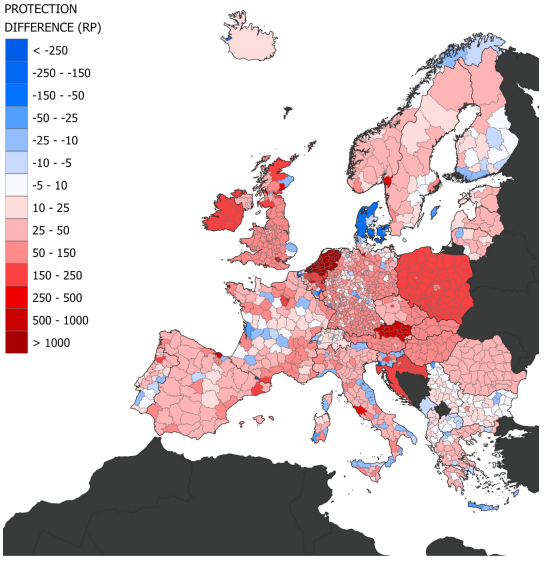

Figure A3. Difference in protection RP between the two flood protection datasets by NUTS3 region (FLOPROS - P25 protections). Red (blue) regions indicate higher (lower) levels of protection in the FLOPROS dataset.

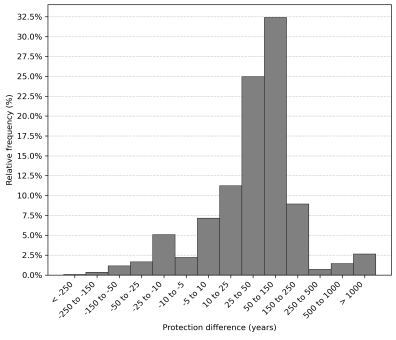

Figure A4. Distribution of the difference between FLOPROS and P25 protections in terms of RP at NUTS3 resolution.

Figures A1 and A2 show the maps and the distribution of flood protections for the two datasets used in this work. FLOPROS protections are higher throughout the continent (with the exception of Denmark and northern Norway), with around 60% of regions being protected from flood with RP lower than 75 years. On the opposite, in the P25 dataset, only about 10% of regions are protection from the 75 year flood. The two flood protection datasets then represent very different estimated of the current flood protections in Europe, which is supported by the results found in this study. In Figures A3 and A4 the map and distribution of the difference between protection RP is also reported: for 70% of NUTS3 regions, the FLOPROS RP estimates are higher by more than 25 years, with some extreme cases like Poland, Austria, Croatia Ireland and the Netherlands, while

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

for only around 12% of regions P25 protections are higher, mainly located in Denmark. These differences have substantial implications for simulated damages, number of events, and country-level risk ranking.

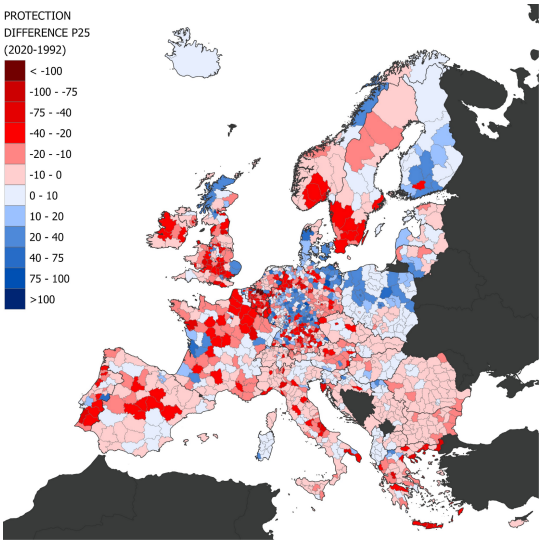

Figure A5. Difference in protections from the P25 dataset between 2020 and 1992. Regions where protections have been estimated to increase are represented in blue, while regions in red have been estimated to have experienced a decrease in protection RP.

The P25 dataset (Paprotny et al., 2025) reports estimated flood protection values from 1950 to 2020 with annual timestep. In Figure A5 the difference between protections in 2020 and in 1992 (the starting year for the analysis) are reported for each NUTS3 region. The predominance of regions in red signifies that, according to the dataset, protections in most european regions have decreased or slightly decreased in the last 3 decades, while significant increases have been estimated only for Finland, Poland, Iceland and, to a lesser extent, parts of Germany, Hungary, France and the UK. Consequently considering the precise annual values reported in the original P25 datasets would have caused even higher flood impacts and number of floods compared to those found in this work, although probably not dramatically since the majority of regions where protections have been estimated to decrease, have done so by less than a RP of 20 years. Due to these reasons then, the dubious likelihood of the phenomena and the necessity to align with the FLOPROS timeframe, only the average protections between 2015 and 2020 were considered, despite impacts were then estimated for floods going back to 1992.

A2 Land Cover Groups

Table A1: LUISA 2018 original land cover classes and land cover group to which each class was assigned to.

Code	Label	Class	Land cover group
1111	High density urban fabric	1	Residential

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Code	Label	Class	Land cover group
1121	Medium density urban fabric	2	Residential
1122	Low density urban fabric	3	Residential
1123	Isolated or very low density urban fabric	4	Residential
1130	Urban vegetation	5	Residential
1210	Industrial or commercial units	6	Industrial, Business
1221	Road and rail networks and associated land	7	Transport
1222	Major stations	8	Transport
1230	Port areas	9	Transport
1241	Airport areas	10	Transport
1242	Airport terminals	11	Transport
1310	Mineral extraction sites	12	Industrial, Business
1320	Dump sites	13	Industrial, Business
1330	Construction sites	14	Industrial, Business
1410	Green urban areas	15	Residential
1421	Sport and leisure green	16	Industrial, Business
1422	Sport and leisure built-up	17	Industrial, Business
2110	Non irrigated arable land	18	Agriculture
2120	Permanently irrigated land	19	Agriculture
2130	Rice fields	20	Agriculture
2210	Vineyards	21	Agriculture
2220	Fruit trees and berry plantations	22	Agriculture
2230	Olive groves	23	Agriculture
2310	Pastures	24	Agriculture
2410	Permanent crops	25	Agriculture
2420	Complex cultivation patterns	26	Agriculture
2430	Land principally occupied by agriculture	27	Agriculture
2440	Agro-forestry areas	28	Agriculture
3110	Broad-leaved forest	29	Other
3120	Coniferous forest	30	Other
3130	Mixed forest	31	Other
3210	Natural grassland	32	Other
3220	Moors and heathland	33	Other
3230	Sclerophyllous vegetation	34	Other

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Code	Label	Class	Land cover group
3240	Transitional woodland shrub	35	Other
3310	Beaches, dunes and sand plains	36	Other
3320	Bare rock	37	Other
3330	Sparsely vegetated areas	38	Other
3340	Burnt areas	39	Other
3350	Glaciers and perpetual snow	40	Other
4000	Wetlands	41	Other
5110	Water courses	42	Other
5120	Water bodies	43	Other
5210	Coastal lagoons	44	Other
5220	Estuaries	45	Other
5230	Sea and ocean	46	Other

A3 Historical Extreme Streamflow Analysis

Being the basis for flood extent simulations, discharge return periods are the basis upon which the work is built. Consequently unrealistic or biased RP estimates would propagate through the entire analysis and affect the resulting flood extents and impact estimates. To evaluate the reliability of the simulated discharge peaks, we assessed the likelihood that the simulated historical peaks could realistically occur within the observation period. Specifically, we estimated the probability that the flood peaks simulated for each river pixel during the 33-year study period could actually have been observed. To do this the individual probabilities of occurrence of a discharge within an RP range (see Section 2.2) were used to generate synthetic realizations of flood occurrence. For each RP class, random realizations of flood series were generated by sampling from a binomial distribution, where the probability parameters of each RP range and the number of trials equal to the length of the historical record (33 years, Equation A1). This procedure was repeated one million times per river pixel, producing one million possible realizations of flood occurrence across all RP classes within a 33-year period.

To assign an overall probability metric to each realization (with length of 33 years), the probabilities of observing the extracted number of events in each RP class were multiplied together (Equation A2).

$$N_i \sim \text{Binomial}(n_{\text{years}}, \tilde{p}_i) \tag{A1}$$

$$L = \prod_{i=1}^{K} \left[\binom{n_{\text{years}}}{N_i} \tilde{p}_i^{N_i} (1 - \tilde{p}_i)^{n_{\text{years}} - N_i} \right] \tag{A2}$$

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

The same approach was then applied to the original observed number and magnitude of simulated flood peaks (in RP terms) at each river pixel. In this way each river pixel was characterized by a historic probability metric and a set of 1 million synthetic probability metrics. Finally, for each pixel, the proportion of synthetic realizations with probability metric lower than the historical metric was computed and interpreted as the likelihood of actually observing the simulated peaks. The results of this analysis are presented in Section B1.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

A4 Flood Peaks Clustering

Algorithm A1 Representation of the algorithm for the clustering of discharge peaks into flood events.

Require: Set of discharge peaks P with spatial coordinates and time
Require: Maximum temporal distance $T_{max} = 90$ days
Require: Initial spatial search radius $d_0 = 30$ km
Require: Maximum spatial search radius $d_{max} = 50$ km
Require: Maximum number of iterations $I_{max} = 50$
Ensure: Catalogue of flood events E

Sort peaks P chronologically
Initialize set of unassigned peaks $U \leftarrow P$
Initialize empty list of flood events E
while U is not empty **do**
    Select earliest peak p_0 from U
    Initialize event $e \leftarrow \{p_0\}$
    Remove p_0 from U
    Compute convex hull H of peaks in e
    Set iteration counter $i \leftarrow 0$
    while $i < I_{max}$ **do**
        Identify peaks in U satisfying:
            temporal distance from $p_0 \leq T_{max}$
            spatial distance from $H \leq d_0$ (first iteration)
            or $\leq d_{max}$ (subsequent iterations)
        if no peaks are found **then**
            Break
        end if
        Add identified peaks to event e
        Remove them from U
        Update convex hull H
        $i \leftarrow i + 1$
    end while
    Compute event start and end time
    Compute event location as centroid of H
    Add event e to list E
end while
return set of flood events E

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

A5 Flood Extent and Impact Calculation

The following section describes more in depth how the flood extent, water depth and impacts of a flood event are calculated.

Let i denote a floodplain pixel and k a river pixel associated to a discharge peak during flood event e. Each river pixel generates a flood hazard tile per RP, providing water depth at every location where this would be > 0, hence identifying the influence area of river pixel k during event e. The water depth at floodplain pixel i generated by river pixel k for return period RP_k is then:

$$h_{i,k}(RP_k) \tag{A3}$$

Let $r(i)$ be the NUTS3 region containing pixel i and $P_{r(i)}$ the flood protection level (expressed as return period) of $r(i)$. Flooding from river pixel k is considered only if the return period of the associated discharge peak exceeds the protection level at location i:

$$\tilde{h}_{i,k} = \begin{cases} h_{i,k}(RP_k), & RP_k > P_{r(i)} \\ 0, & RP_k \leq P_{r(i)} \end{cases} \tag{A4}$$

The final water depth at pixel i for event e is then defined as the maximum depth across all contributing river pixels:

$$h_i^{(e)} = \max_k \tilde{h}_{i,k} \tag{A5}$$

The river pixel responsible for the inundation is

$$k_i^{(e)} = \arg h_i^{(e)} \tag{A6}$$

with associated return period

$$RP_i^{(e)} = RP_{k_i^{(e)}} \tag{A7}$$

Let now L_i denote the land-cover class of pixel i, and let $D_{r,L}(h)$ be the damage associated with the depth-damage function $D_{r,L}$ and water depth h, with r representing again the NUTS3 region. The economic damage at pixel i is computed then as

$$d_i = D_{r(i),L_i}\left(h_i^{(e)}\right), \tag{A8}$$

while the population of the pixel is simply considered in its entirety once $h_i^{(e)} > 0$. Some examples of the results of this procedure are reported in Figures A6 and A7.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

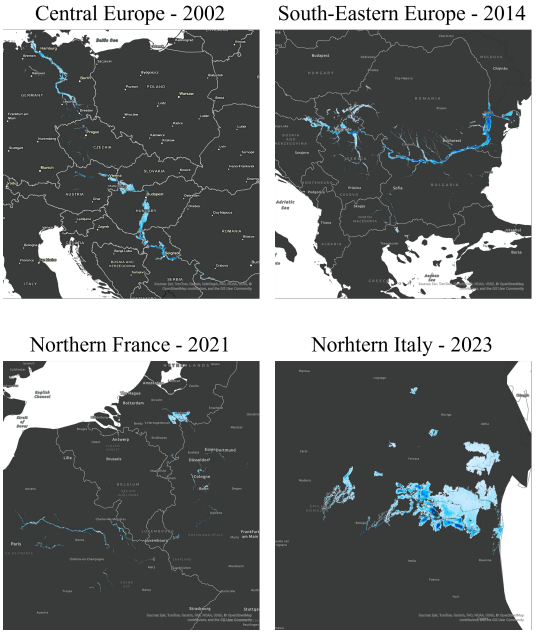

Figure A6. Examples of flooded area and depth for 4 major flood events.

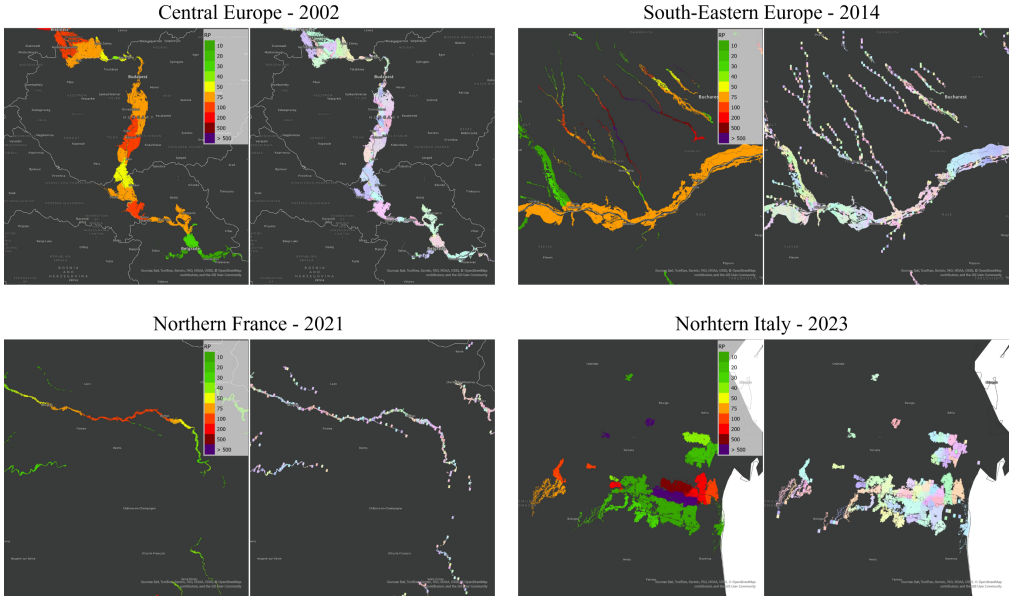

Figure A7. Example of data concerning the river pixel causing the maximum flood depth and the corresponding RP for 4 major flood events. The right panel of each event shows with the same color each river pixel and the associated final flooded area.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Figure 10 represents the maximum number of river pixels affecting each floodplain pixel and the maximum flooding distance of each river pixel respectively. For large rivers, each point of the 500 y RP floodplain could potentially be flooded by several river pixels, up to around 150 in the cases on the Danube, Po' and Rhine. The lower course of the Danube in particular present a situation in which flooding could come from a large number of pixels, potentially located upstream by more than 200 km. Despite a lower number of affecting river pixels, a similar situation is found for all major European rivers, especially on reaches located in plains. This fact justifies the large extents of simulated floods originating from these rivers, even when observed extents wouldn't verify it.

A6 Event matching

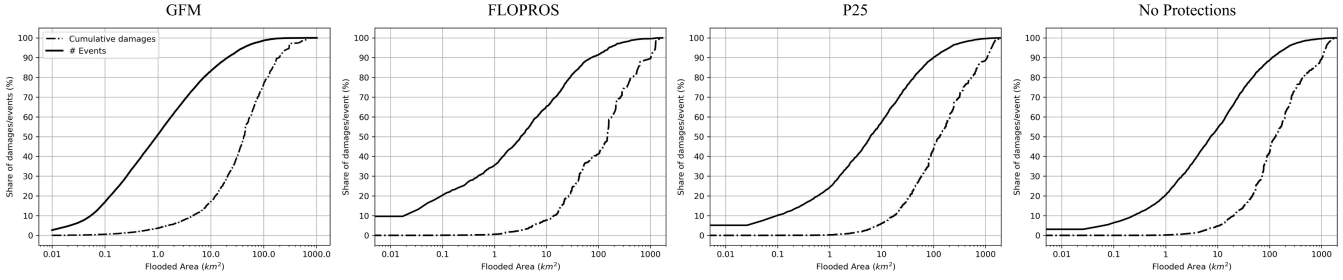

Figure A8. Cumulative number of events and damages from the GFM and simulations (2015-2024) with the three protection settings, sorted by increasing flooded area.

Figure A8 shows the cumulative distribution of GFM events and simulations and damages, sorted by flooded area size. Approximately 51% of all GFM events, after being cut on hydrological basins, present a flooded area lower than 1 km^2, but these account only for 4% of total damages. In the case of simulations, the share of events with flooded area lower than the threshold is between 20 and 35 %, with share of damages lower than 1%. We can also note that decreasing the overall levels of protections leads shares of small events to increase, due to a decreasing of the total number of events, meaning that introducing higher protections mainly limits the occurrence of large floods, while smaller floods probably deriving from small rivers still get simulated, due to higher RP. This is coherent with the expectation that protections are more often built on bigger rivers rather than small rivers.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Algorithm A2 Representation of the algorithm for matching simulated flood events and GFM observed events.

Require: Set of simulated events S
Require: Set of GFM events G
Require: Temporal tolerance Δt (weeks)
Require: Minimum temporal overlap threshold τ
Require: Spatial aggregation level (e.g. NUTS2)
Require: Influence areas linking river pixels to floodplain cells
Ensure: Set of matched event groups M

Combine S and G into a single list of events E
Assign a unique identifier to each event
Initialize graph H with one node per event in E
for each simulated event $s \in S$ **do**
    Extend temporal window of s by $\pm\Delta t$
    for each GFM event $g \in G$ **do**
        Compute temporal overlap duration: $overlap = \min(end_s, end_g) - \max(start_s, start_g)$
        Compute normalized overlap: $overlap\_ratio = \frac{overlap}{\min(duration_s, duration_g)}$
        Determine spatial proximity:
            events share at least one NUTS region or basin
        if $overlap\_ratio > \tau$ AND spatial proximity exists **then**
            Identify river pixels influencing s
            Identify river pixels influencing g
            if the two sets of river pixels intersect **then**
                Add edge between s and g in graph H
            end if
        end if
    end for
end for
Identify connected components of graph H
for each component C **do**
    if C contains both simulated and GFM events **then**
        Aggregate events into a matched event group
        Compute combined time span and affected regions
    end if
end for
Merge groups that share common simulation identifiers
return matched event groups M

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Appendix B: Supplementary Results

B1 Historic Extreme Streamflow Analysis

The simulation of historical floods fundamentally relies on the accuracy of streamflow return periods (RPs). Unrealistic or biased RPs would propagate through the flood hazard maps, affecting the identification of flooded areas and associated damages. Figure B1 evaluates the plausibility of the simulated flood peaks: for 42% of river pixels, the likelihood of observing the simulated RPs exceeded 75%, and for 87% of pixels, it exceeded 25%. Only 4.1% of pixels had a likelihood below 10% (Figure B2); and around 40% of these low-likelihood pixels corresponded to small upstream areas (<450 km^2, Figure B3). Overall, these findings indicate that the simulated flood peaks provide a realistic representation for large-scale risk assessment, although direct validation with observed discharge data would be required to further confirm accuracy.

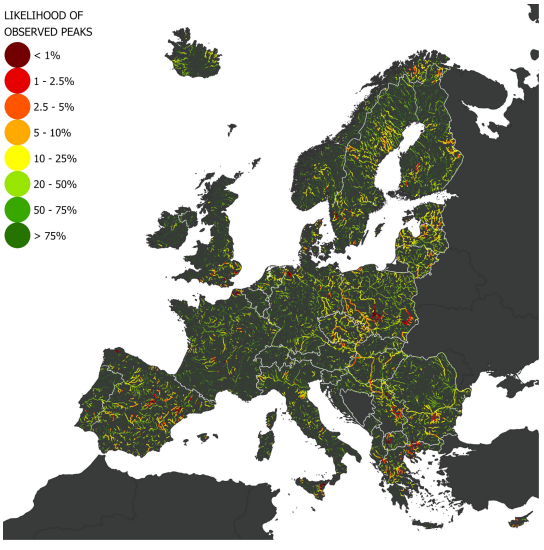

Figure B1. Likelihood of occurrence of the simulated historic flood peaks

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

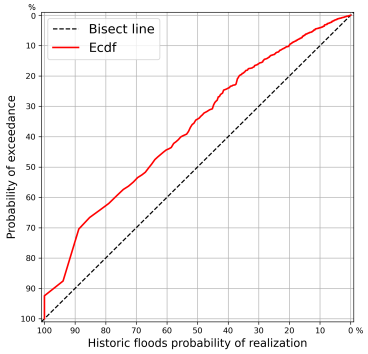

Figure B2. Likelihood probability and historical ecdf of all river pixels.

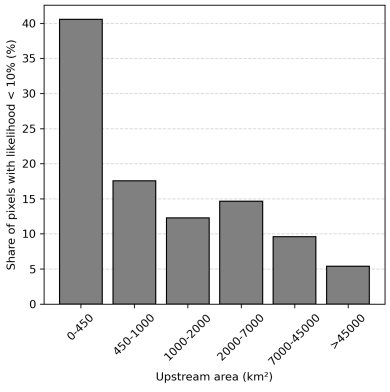

Figure B3. Share of pixels with likelihood lower than 10% divided by upstream area class

B2 Historic Impact Assessment

Figure B4 shows the number and total extent of flooded areas simulated through the use of the three protection datasets. Once again, FLOPROS causes a much smaller number and total extent of floods especially in Western Europe and Scandinavia, while P25, compared to the case with no protections, doesn't lead to significant decreases for the whole of Eastern Europe and the Mediterranean area. Particularly and repeatedly impacted are the areas around the Danube and Tisza rivers, between Slovakia, Hungary, Romania and Serbia, as well as near the Po' river outlet. Considering the lower P25 protections rather than FLOPROS would lead to significantly more flooding along the Rhone, Elbe, Guadalquivir, Tago and Vistula rivers, among others.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

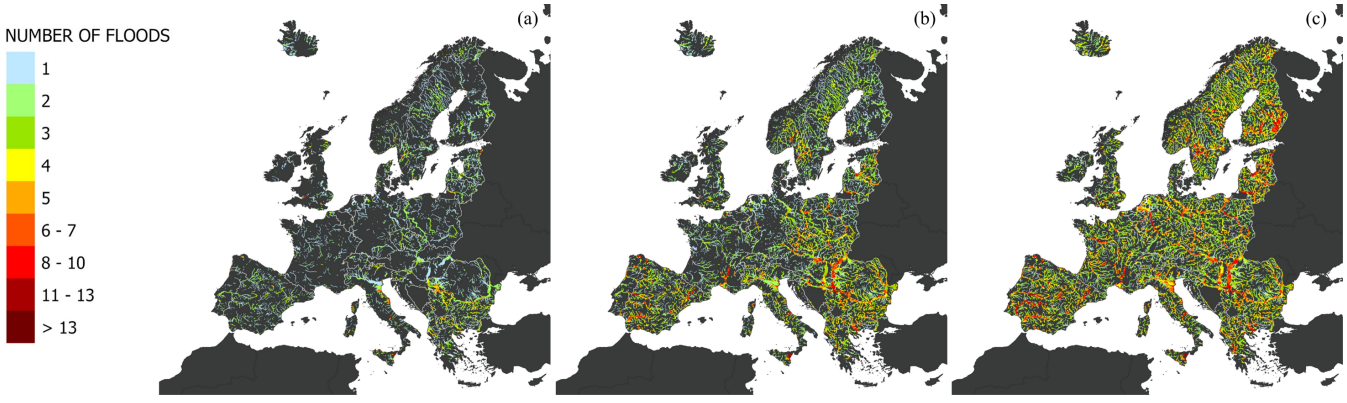

Figure B4. Total number of and extent of floods from 1992 to 2024 with FLOPROS protections (a), P25 protections (b) and No Protections (c).

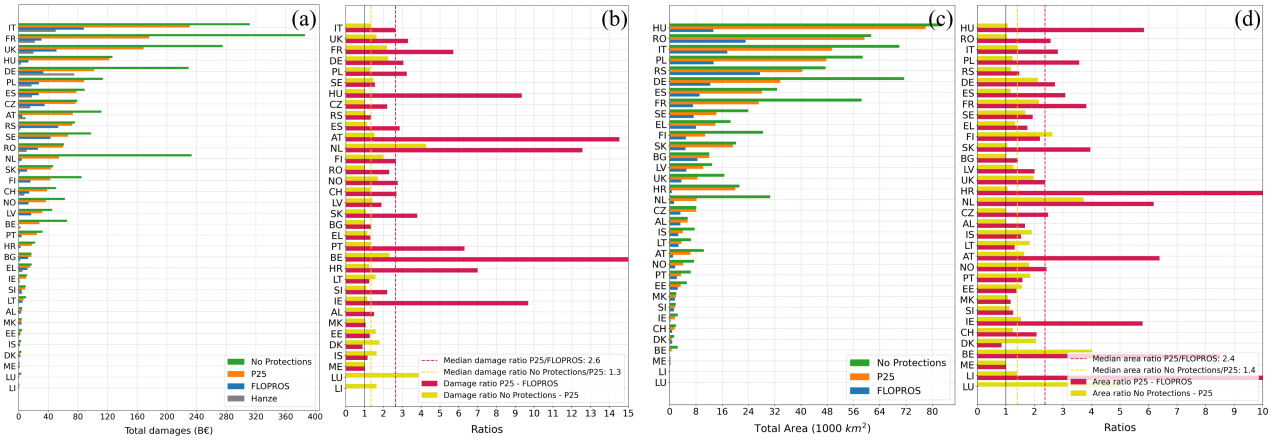

Figure B5. Comparison between the total damages (a) and total flooded areas (c) obtained through the use of the three protection datasets and the respective ratios (b,d) for all the analysed countries.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

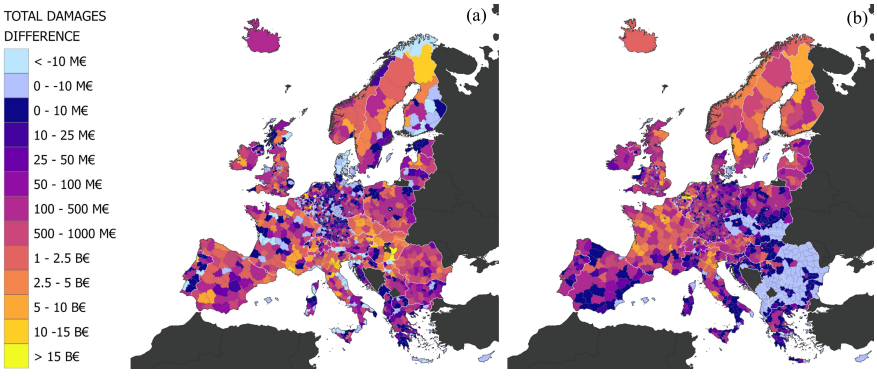

Figure B6. Difference of total damages between P25 and FLOPROS protections (a) and between No Protections and P25 (b) by NUTS3 region

Figure B5 shows the comparison between all three protection datasets of total damages and flooded areas for all analysed countries, ranked by average rank across the datasets. Using FLOPROS dataset rather than P25 causes in median an increase in flooded area by 2.4 times, while not using protection a further increase by 1.4 folds. Overall then FLOPROS protection cause about 3.4 times more damages and flooded areas than a case without flood protections, with significantly higher increases in Belgium, Austria, the Netherlands, Hungary and Ireland.

B3 Comparison of observed, simulated and reported events

Matches, misses and false alarms

In Figure B7 the location of false alarms is shown. With FLOPROS protections, false alarms are mainly located along minor rivers and for relatively short reaches in Spain, Romania, Serbia, Greece, Latvia and Lithuania. With P25 protections though false alarms on some major rivers start to emerge, mainly along the, Tisza, Danube and Morava, as well as between the Southern portion of the border between Sweden and Norway, where many lakes are located, and in Iceland. Moving to the case without protections intensifies the clustering of false alarms around these regions and across the whole continent.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

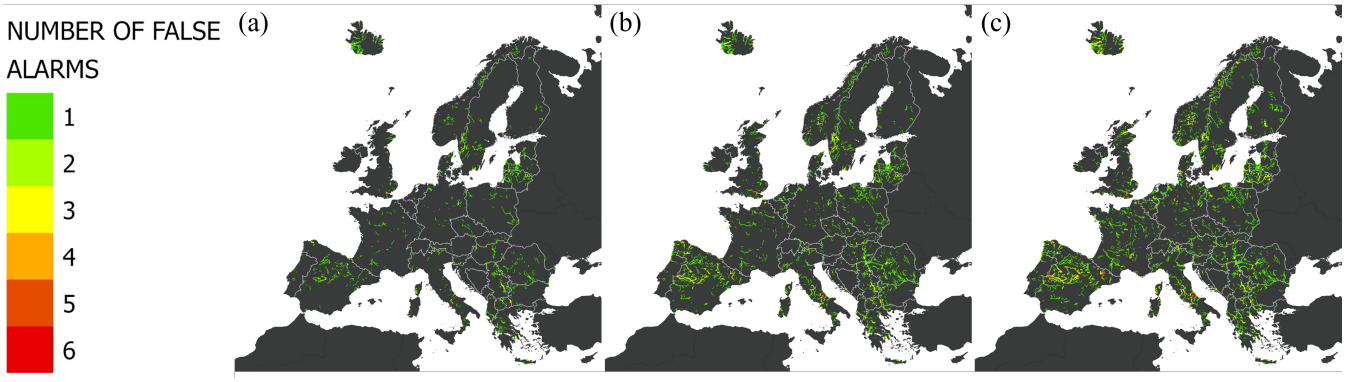

Figure B7. Number, location and extent of false alarms for the FLOPROS (a), P25 (b) and No protection (c) settings.

Concerning the location of misses, a large share are located in Finland and Swedish Lapland, England and Western France. Concerning the misses caused by too high protections, for both FLOPROS and P25 they are located in Ireland, England, both Eastern and Western France and Central and Southern Finland. The cases of the Marne, Charente and Saone, in France, Shannon and Suck, in Ireland and Severn in England are an indication and example of the overestimation of protections by both datasets. These are areas around mainly middle sized rivers where floods have clearly been observed multiple times (at least 3 times in most of these cases) but floods were not simulated not reported, despite significant discharges had indeed been computed. The large number and area of misses in Finland and Scandinavia in general instead can be indicative of a different issue. The presence of these areas across all misses types and the complete absence of impacts in Hanze for Finland, Sweden and Norway in their entirety suggests two distinct problems. On the one hand GFM observations in these areas could be actually largely considered false alarms, due to large presence of snow and wet soil throughout the year that might have escaped the masking process (Betterle and Salamon, 2025) and which could in turn cause matches with simulated events with significant RP in the same period, on the other though, the lack of impacts in the Hanze dataset for these areas can also be caused by the large population centers or human activity in its entirety, which would of course hinder the extraction of flood events by news or governmental sources. Scandinavia then presents even more challenges than the rest of Europe concerning the actual impacts of historic floods. Moving to misses caused by the lack of flood hazard maps (RP between 2.5 and 10 years), they are again located on the same rivers previously identified in the case of France, but with the addition of the Loire, at its outlet. In this case a significant number and extent of misses can be also more widely identified across the continent, especially along the floodplain of the Weser, Tisza, Elbe, Rhine, Sava and Danube. A similar trend continues when analyzing the location of misses caused by the complete lack of significant return period, with an even greater distribution but much higher presence of small rivers and short reaches.

These results highlight likely structural problems for certain rivers (like in the cases of France and Ireland), where discharge RP are lower than what they should theoretically be and it also likely that protection estimates are too high, however the two issues reinforce each other. In addition areas like Scandinavia are particularly challenging due to limitations of all datasets: the

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

large exclusion masks and presence of water look-alikes in the GFM case, the sparseness of stations and calibration data in the case of simulations and the probable lack of reports and governmental sources for floods in large natural areas with no or little human activity in the case of Hanze.

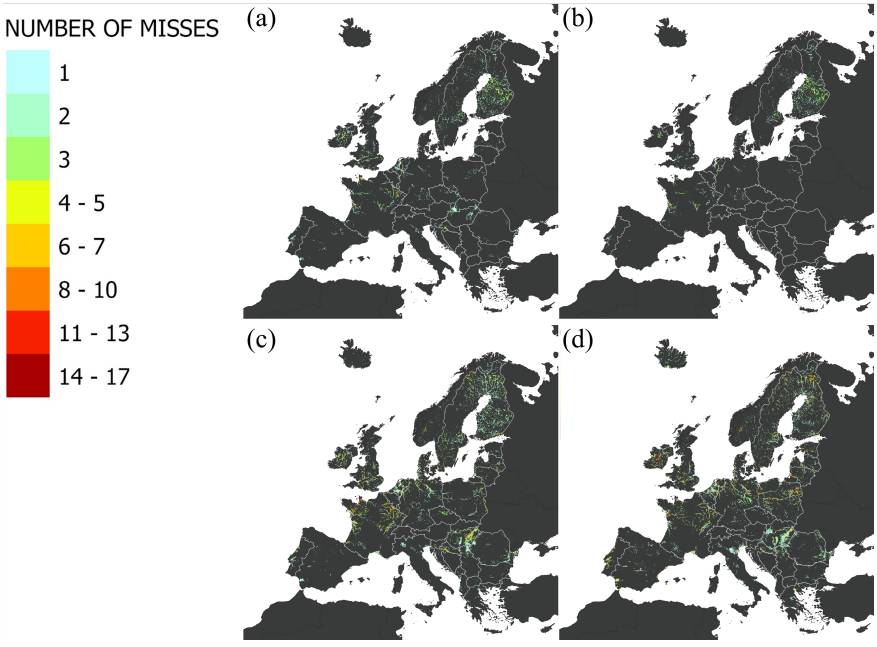

Figure B8. Number, location and extent of misses caused by too high protections for FLOPROS (a) and P25 (b) and misses caused by RP between 10 and 2.5 years (c) and RP lower than 2.5 (d).

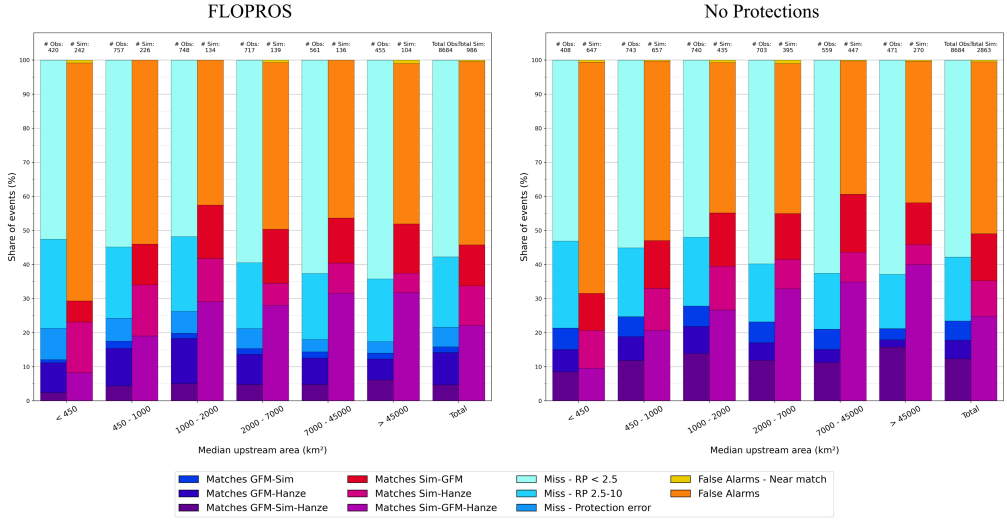

Figure B9. Shares of matches, misses and false alarms by type and catchment size for FLOPROS and No Protections setting.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Matched flood events

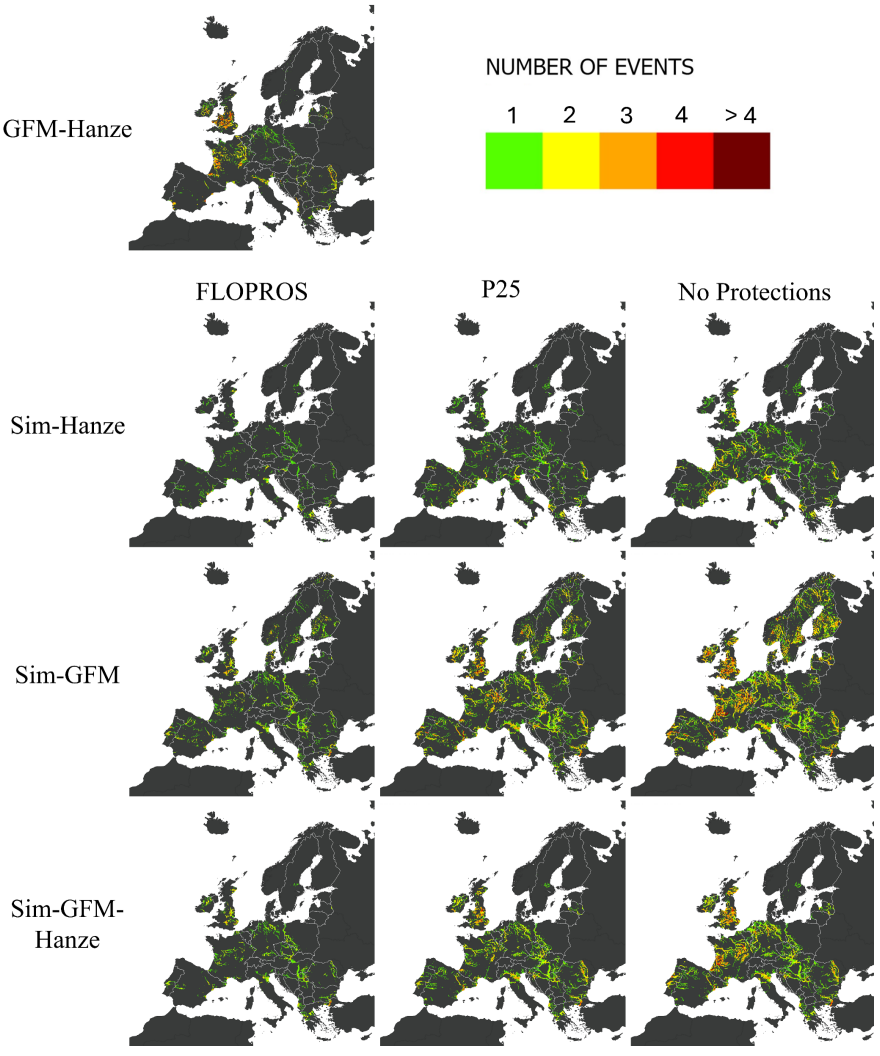

Figure B10. Location, extent and number of matched events between any two dataset combination, with the different protection settings. The extent and number of events in the Sim-GFM and Sim-GFM-Hanze cases account for both simulated and observed flood extents.

B4 Analysis of Matched Events

Simulated damage bias

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

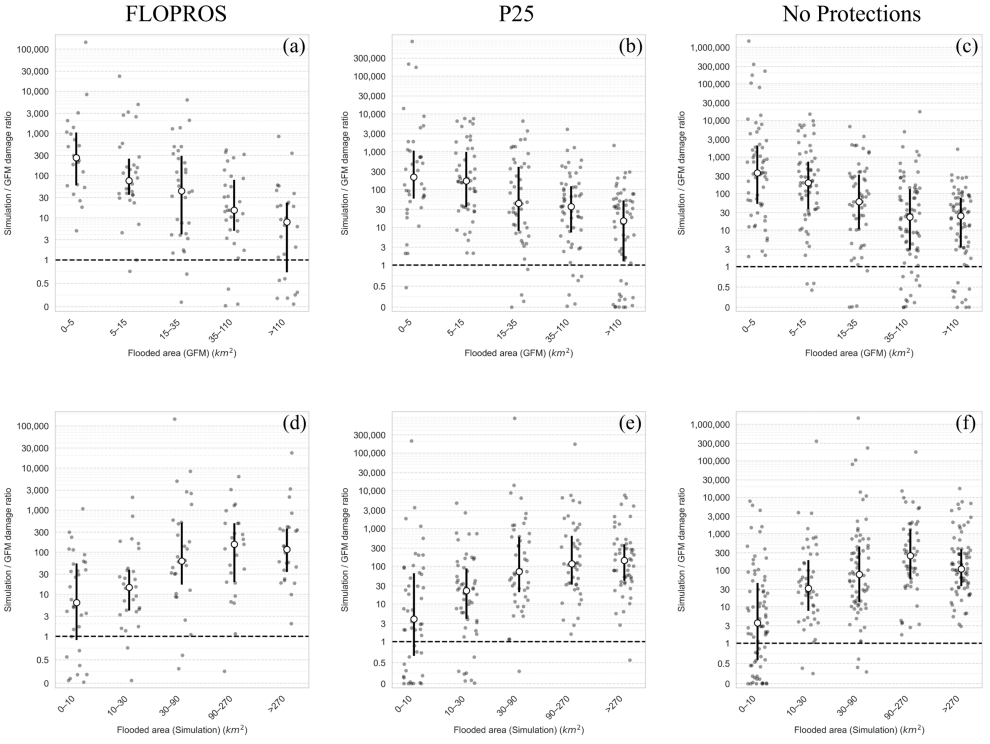

Figure B11. Bias of estimated damages from simulations with respect to estimated damages from GFM for FLOPROS (a,d), P25 (b,e) and No protections (c,f), divided by classes of observed flooded area (a,b,c) and simulated flooded area (d,e,f).

Overlap area between simulated and observed events

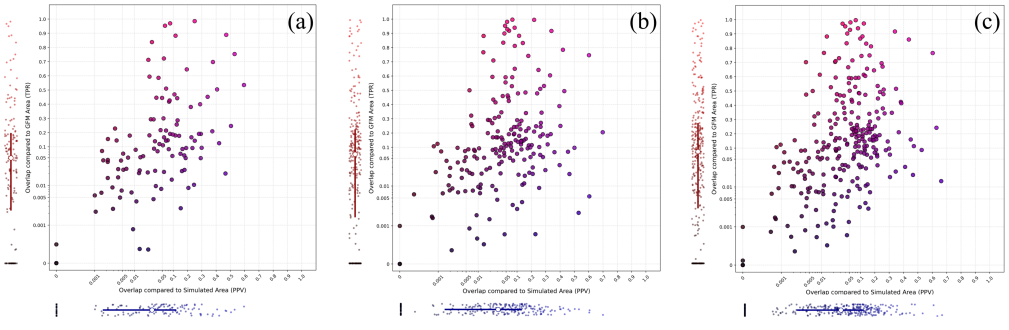

Figure B12. Scatterplot of the share of overlap area compared to the total observed and total simulated extents, for FLOPROS (a), P25 (b) and No protections (c).

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

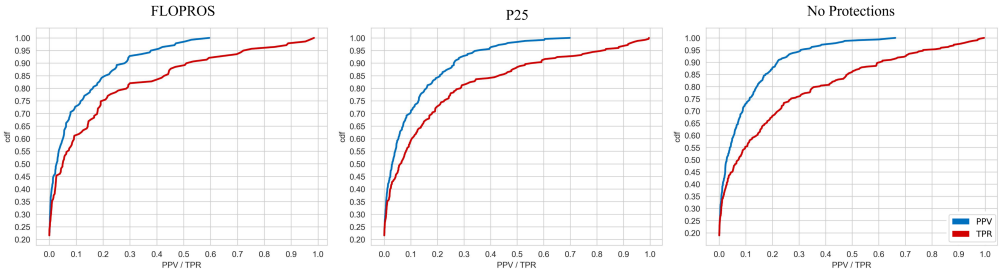

Figure B13. Cumulative distribution of PPV and TPR for the three protection settings

Table B1. Spearman correlation and 5-95% confidence intervals between bias in simulated flooded area and total flooded area (identified by simulations and by observations individually) for matched events between simulations, observations and Hanze.

Total flooded area	FLOPROS	P25	No Protections
Simulation	0.53 [0.34, 0.68]	0.48 [0.32, 0.61]	0.4 [0.22, 0.56]
GFM	-0.54 [-0.69, -0.34]	-0.43 [-0.58, -0.27]	-0.42 [-0.58, -0.25]

Table B2. Spearman correlation coefficients and 5–95% confidence intervals between PPV (and TPR) and total flooded area for matched events between simulations, observations and Hanze.

	FLOPROS		P25		No Protections	
Total flooded area	PPV	TPR	PPV	TPR	PPV	TPR
Simulation	0.17 [-0.07, 0.39]	0.43 [0.20, 0.63]	0.24 [0.05, 0.42]	0.51 [0.35, 0.65]	0.34 [0.16, 0.51]	0.53 [0.37, 0.67]
GFM	0.38 [0.18, 0.56]	-0.05 [-0.28, 0.19]	0.56 [0.43, 0.67]	0.11 [-0.08, 0.30]	0.61 [0.48, 0.71]	0.16 [-0.02, 0.34]

Water depth and affected land cover classes

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

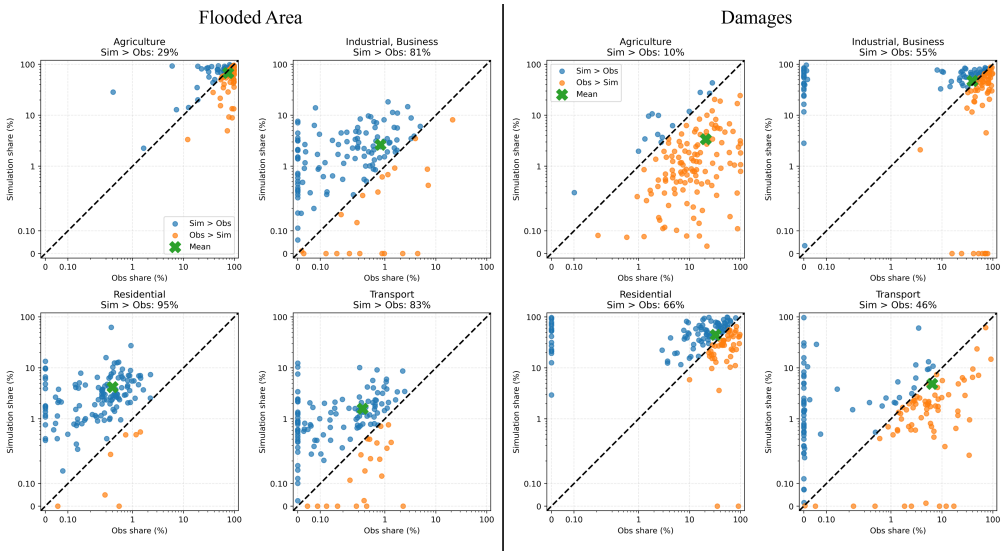

Figure B14. Scatterplot of flooded area and damage shares, FLOPROS protections.

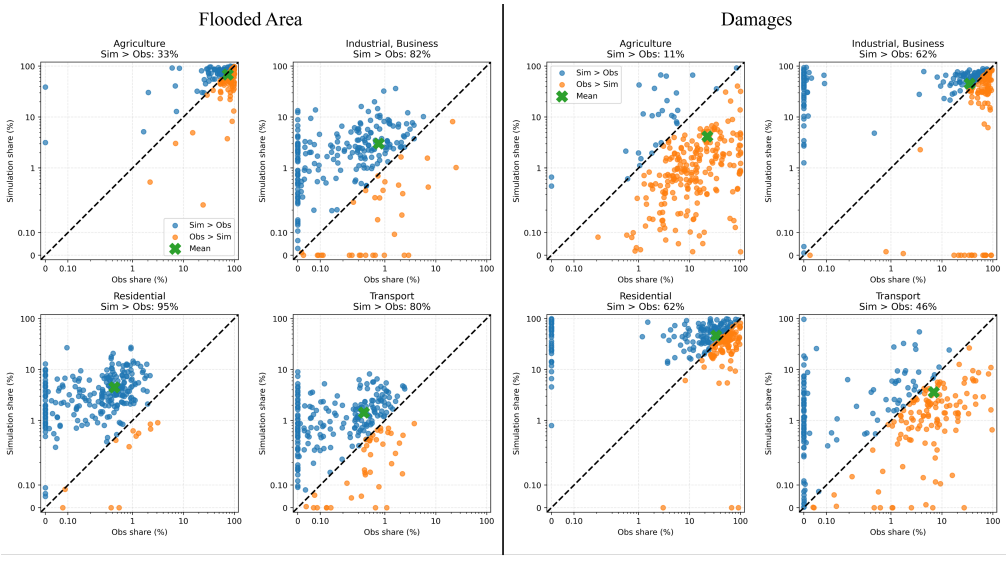

Figure B15. Scatterplot of flooded area and damage shares, P25 protections.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

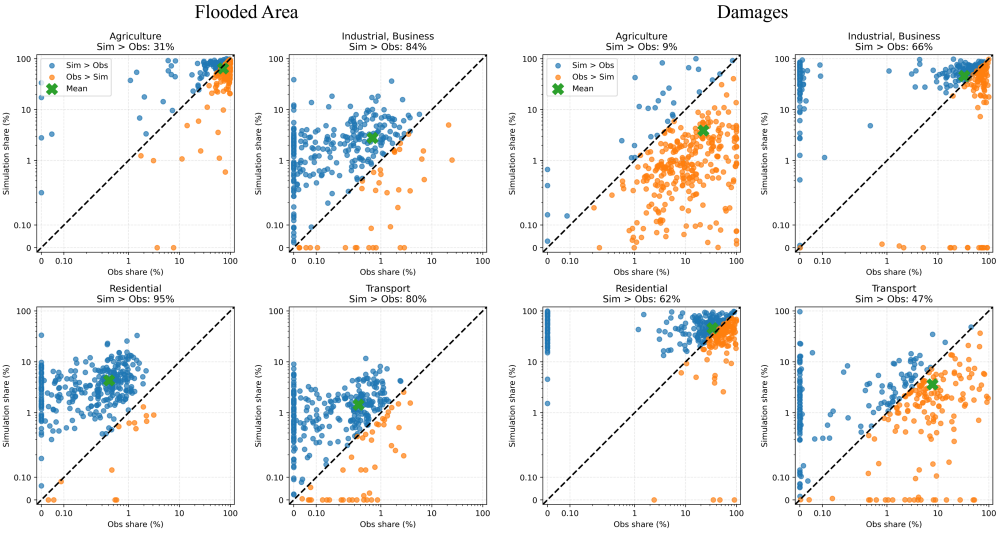

Figure B16. Scatterplot of flooded area and damage shares, No Protections.

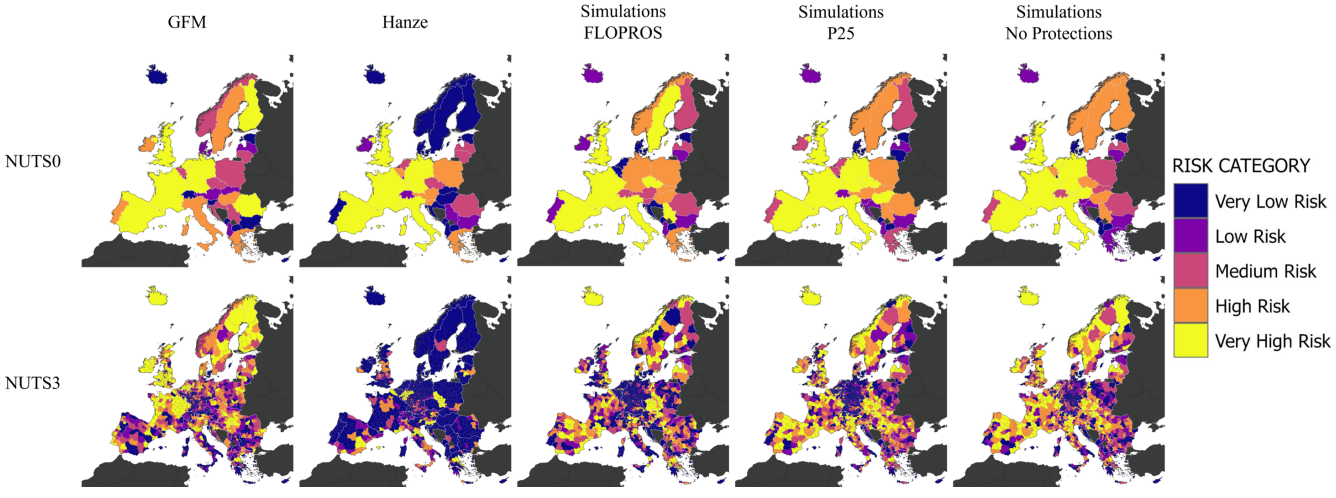

Figure B17. Risk categories at NUTS0 and NUTS3 levels for reported impacts, GFM events and simulations.

B5 Sensitivity analysis

The following section provides a sensitivity analysis for the main parameters used in this work. The sensitivity analyses were conducted for the No Protection setting, since this would be the protection setting where the most differences could be experienced with varying parameter values.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

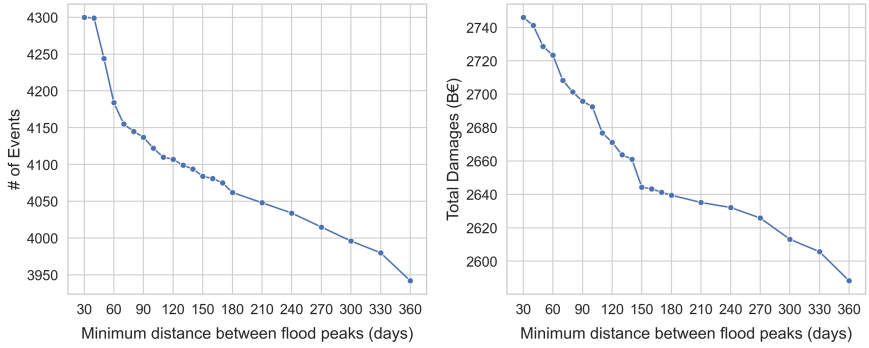

Figure B18. Sensitivity analysis to the minimum number of days between flood peaks. The left panel reports the number of identified flood events with different parameter values, while on the right the total simulated damages are shown.

In Figure B18 the number of floods and the total damages simulated under the No Protection scenarios and different thresholds for the minimum separation between flood peaks occurring on the same river pixel. As visible from the left panel using a minimum separation lower than 90 days, and in particular lower than 70 days, would significantly increase the number of simulated events since multiple floods would be simulated more than once, accounting for different peaks. As a consequence also the total damages would increase: with a minimum distance between peaks of 30 days, the damages would be around 2% higher than with a minimum separation of 90 days, while the number of events would be around 3.8% higher. Increasing the minimum separation on the other hand would further reduce the damages and number of events, up to a reduction of 4.7% and 3.7% respectively with a single peak per year. The choice of the minimum threshold of 90 days selected in this work then is a compromise between the need to avoid counting more than once the same flood event or separate events happening in a very short time-span and the need to still consider the possibility of having multiple floods per year on the same river and in the same region, allowing for both recovery and multiple loss events.

Table B3. Sensitivity analysis on the minimum damage threshold for matching Simulated and Hanze events.

Minimum damage threshold	Simulations matched	# Simulations	Hanze (Total)	Hanze (With impacts)
0	716 (17.67%)	4053	45.96%	54.01%
0.1	681 (19.03 %)	3578	45.08%	53.71%
1	642 (20.44 %)	3141	44.44%	53.25%

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table B4. Sensitivity analysis on the tolerance (in weeks) for matching simulated and Hanze events.

Tolerance (weeks)	Simulations matched	Hanze (Total)	Hanze (With impacts)
1	681 (19.03 %)	45.08 %	53.71 %
2	735 (20.24 %)	47.48 %	55.82 %
3	773 (21.28 %)	49.08 %	57.19 %
4	808 (22.25 %)	51.00 %	58.70 %
5	840 (23.13 %)	51.96 %	59.30 %
6	878 (24.17 %)	53.00 %	59.76 %
7	911 (25.08 %)	54.44 %	61.57 %
8	949 (26.13 %)	55.16 %	62.33 %
10	1009 (27.78 %)	57.15 %	63.69 %
12	1071 (29.49 %)	58.75 %	64.90 %
16	1186 (32.65 %)	62.35 %	68.23 %
20	1285 (35.38 %)	65.39 %	70.95 %

Tables B3 and B4 show the total number of simulations and the share of matches with Hanze obtained using different parameters concerning the minimum event damage and the tolerance in weeks for matching the events. Eliminating the threshold on the minimum damages of the simulations when matching Hanze would increase the number of simulation matches by 35 and the share of matches in Hanze by 0.3%, but at the same time the share of simulation matches would decrease by 1.4%, indicating that the majority of the events that would be introduced by not considering the threshold would in fact not be matched. Increasing the threshold to 1 million € would have a similar overall effect, with a reduction in the number of simulation matches of 39 but an increase in their share of 1.4 %. Overall then these differences are minimal and do not affect the broader message. Increasing the semi-amplitude of the tolerance window for matching the events in time to 5 weeks instead of 1, would increase the share of matches of the simulations by 5% and that of Hanze by around 7%, while a further increase to 10 weeks would bring an additional increase in matches by 4.5 % and 5.2 % respectively. It has to be expected that more relaxed constraints would increase the number of matches, however it must be noted that with even a semi-amplitude of 20 weeks (which would mean searching for events affecting the same NUTS2 region in a period from 20 weeks before the Hanze starting date to 20 weeks after the end date) the share of matches would not go above 36% for the simulations and 66% for Hanze, still indicating a low agreement between the datasets despite the relaxed constraint and consequently the large probability of matching events that should not actually be considered the same flood.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Table B5. Sensitivity analysis on the minimum flooded area threshold for matching Simulated and GFM events.

Minimum flooded area threshold	Sim matched	# Sim events	GFM matched	# GFM events
0.01	1469 (41.5 %)	3539	2205 (12.5 %)	17674
0.1	1389 (41.2 %)	3373	2084 (14 %)	14913
1	1100 (38.4 %)	2863	1546 (17.8 %)	8684
10	555 (33.7 %)	1647	603 (20.3 %)	2967

Table B6. Sensitivity analysis on weeks of tolerance for matching Simulated and GFM events.

Tolerance (weeks)	Sim matched	GFM matched
1	38.4%	17.8%
2	40.3%	19.4%
3	41.6%	20.9%

Tables B5 and B6 instead show the sensitivity to the parameters of matches between simulations and GFM events. The considerations are similar to those for simulations and Hanze. Relaxing the tolerance windows for matching events would not increase significantly the share of matches but would increase the chance of matching unrelated events, introducing further uncertainty in the matching procedure without actual benefits of fundamental changes in the results. Considering a different threshold for the minimum flooded area needed in order to consider an event from both sources again shows that, despite the results being sensitive to these parameters, the overall message doesn't change. The overall number of events would increase by reducing the threshold from 1 km^2 to 0.1 or event 0.01 km^2, however the share of matches would decrease. It is evident how this threshold has a bigger impact on the GFM events than the simulations, given the large share of events with small areas. In fact a threshold of 0.1 km^2 increases the overall number of GFM events to be considered in the matching procedure by more than 70% and the absolute number of GFM matches by 538. However one must consider that these new events that get introduced in the analysis are exceptionally small to represent true river floods and it is much more likely for them to represent other phenomena or minor areas of larger floods. This is confirmed by the fact that the number of simulation matches would not increase so drastically. Moving from a threshold of 1 to 10 km^2 on the other hand would halve the number of matches in both the simulations and GFM, but this would have different effects on the datasets, in fact the share of simulated matches would decrease by 4.7% while that of the GFM would increase by 2.5%, confirming once again the validity of the results highlighted in this work, namely that the reliability of floods observed from satellites is larger when these floods are also large, while the same is not necessarily true for the simulations.

Author contributions.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

L. Scarpellini: Conceptualization (lead), Methodology (equal), Software, Formal analysis, Investigation, Visualization, Writing - Original Draft (lead). **A. Ficchì:** Conceptualization (equal), Methodology (equal), Supervision (equal), Writing - Review & Editing (lead). **C. D'Angelo:** Conceptualization (equal), Methodology (equal), Supervision (equal), Writing - Review & Editing (supporting). **A. Betterle:** Conceptualization (equal), Methodology (equal), Supervision (equal), Writing - Review & Editing (supporting). **A. Castelletti:** , Supervision (supporting), Writing - Review & Editing (supporting), Resources (equal), Funding acquisition, Project administration. **P. Salamon:** Conceptualization (equal), Methodology (supporting), Supervision (supporting), Writing - Review & Editing (supporting), Resources (equal), Project administration.

Competing interests.

The authors have no conflicts of interest to disclose.

Acknowledgements. Lorenzo Scarpellini led this article while attending the PhD programme in Sustainable Development and Climate Change at the University School for Advanced Studies IUSS Pavia, Cycle 39, with the support of a scholarship co-financed by the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4, Component 1 - Investment 4.1.

Andrea Ficchì and Andrea Castelletti were supported under the SOS-WATER project and funded by the European Union's Horizon EUROPE Research and Innovation Programme call [HORIZON-CL6-2021-CLIMATE-01-Grant Agreement Number 101059264].

We thank colleagues from the CEMS-Floods teams at the JRC and ECMWF for their work on the Copernicus Emergency Management Service (CEMS) EFAS v5.0.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

References

Aerts, J. P., Uhlemann-Elmer, S., Eilander, D., and Ward, P. J.: Comparison of estimates of global flood models for flood hazard and exposed gross domestic product: a China case study, Natural Hazards and Earth System Sciences, 20, 3245–3260, 2020.

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P., and Feyen, L.: Advances in pan-European flood hazard mapping, Hydrological processes, 28, 4067–4077, 2014.

Bates, P. D., Horritt, M. S., and Fewtrell, T. J.: A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling, Journal of hydrology, 387, 33–45, 2010.

Bauer-Marschallinger, B., Cao, S., Tupas, M. E., Roth, F., Navacchi, C., Melzer, T., Freeman, V., and Wagner, W.: Satellite-based flood mapping through Bayesian inference from a sentinel-1 SAR datacube, Remote Sensing, 14, 3673, 2022.

Baugh, C., Colonese, J., D'Angelo, C., Dottori, F., Neal, J., Prudhomme, C., and Salamon, P.: River flood hazard maps for Europe and the Mediterranean Basin region, https://doi.org/10.2905/1D128B6C-A4EE-4858-9E34-6210707F3C81, 2024.

Ben-David, A.: Comparison of classification accuracy using Cohen's Weighted Kappa, Expert Systems with Applications, 34, 825–832, 2008.

Bernhofen, M. V., Whyman, C., Trigg, M. A., Sleigh, P. A., Smith, A. M., Sampson, C. C., Yamazaki, D., Ward, P. J., Rudari, R., Pappenberger, F., et al.: A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique, Environmental Research Letters, 13, 104 007, 2018.

Betterle, A. and Salamon, P.: Water depth estimate and flood extent enhancement for satellite-based inundation maps, Natural Hazards and Earth System Sciences, 24, 2817–2836, 2024.

Betterle, A. and Salamon, P.: Satellite-derived flood depth maps for Europe, https://doi.org/10.2905/0bc96690-b89c-4909-9166-c2c322a20130, 2025.

Birant, D. and Kut, A.: ST-DBSCAN: An algorithm for clustering spatial–temporal data, Data & knowledge engineering, 60, 208–221, 2007.

Boulange, J., Hirabayashi, Y., Rimba, A. B., and Modi, P.: A new generation of global flood protection database, International Journal of Disaster Risk Reduction, p. 105823, 2025.

Burek, P. A., Van der Knijff, J., and De Roo, A.: LISFLOOD-distributed water balance and flood simulation model-revised user manual 2013, Publications Office of the European Union, Luxembourg, 2013, 2013.

Choné, G., Biron, P. M., Buffin-Bélanger, T., Mazgareanu, I., Neal, J. C., and Sampson, C. C.: An assessment of large-scale flood modelling based on LiDAR data, Hydrological Processes, 35, e14 333, 2021.

Coccia, G., Ceresa, P., Bussi, G., Denaro, S., Bazzurro, P., Martina, M., Fagà, E., Avelar, C., Ordaz, M., Huerta, B., et al.: Large-scale flood risk assessment in data scarce areas: an application to Central Asia, Natural Hazards and Earth System Sciences Discussions, 2023, 1–33, 2023.

Cohen, J.: A coefficient of agreement for nominal scales, Educational and psychological measurement, 20, 37–46, 1960.

Collalti, D.: The Economic Dynamics After a Flood: Evidence from Satellite Data: D. Collalti, Environmental and Resource Economics, 87, 2401–2428, 2024.

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L.: Development and evaluation of a framework for global flood hazard mapping, Advances in water resources, 94, 87–102, 2016.

Dottori, F., Kalas, M., Salamon, P., Bianchi, A., Alfieri, L., and Feyen, L.: An operational procedure for rapid flood risk assessment in Europe, Natural Hazards and Earth System Sciences, 17, 1111–1126, 2017.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Dottori, F., Alfieri, L., Bianchi, A., Skoien, J., and Salamon, P.: A new dataset of river flood hazard maps for Europe and the Mediterranean Basin, Earth System Science Data, 14, 1549–1569, 2022.

Dottori, F., Mentaschi, L., Bianchi, A., Alfieri, L., and Feyen, L.: Cost-effective adaptation strategies to rising river flood risk in Europe, Nature Climate Change, 13, 196–202, 2023.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise, in: kdd, vol. 96, pp. 226–231, 1996.

European Environment Agency: Economic losses from weather- and climate-related extremes in Europe, https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related, accessed: 2025-10-14, 2025.

Eurostat: Regions in the European Union — Nomenclature of territorial units for statistics — NUTS 2016/EU-28, https://doi.org/10.2785/475524, 2018.

Eurostat: Gross domestic product (GDP) at current market prices by NUTS 3 region, https://doi.org/10.2908/NAMA_10R_3GDP, 2025.

FAO: Hydrological basins in Europe, https://data.apps.fao.org/catalog/dataset/1849e279-67bd-4e6f-a789-9918925a11a1, 2021.

Huizinga, J., De Moel, H., and Szewczyk, W.: Global flood depth-damage functions: Methodology and the database with guidelines, https://doi.org/10.2760/16510, 2017.

Johnson, T. G., Leandro, J., and Ahadzie, D. K.: Quantifying hazard resilience by modeling infrastructure recovery as a resource-constrained project scheduling problem, Natural Hazards and Earth System Sciences, 24, 2285–2302, 2024.

Joint Research Centre of the European Commission: Disaster losses and damages Dashboard, https://drmkc.jrc.ec.europa.eu/risk-data-hub/#/losses/dashboard, accessed: 2026-02-18, 2026.

Khanh, D. N., Tsumura, Y., Sasaki, O., Shiraishi, K., Akimoto, D., Yamazaki, D., Zhao, G., and Hirabayashi, Y.: Mapping the world's river levees: A hyper-resolution levee database based on digital elevation models, Geophysical Research Letters, 52, e2024GL114 121, 2025.

Matgen, P., Martinis, S., Wagner, W., Freeman, V., Zeil, P., and McCormick, N.: Feasibility assessment of an automated, global, satellite-based flood monitoring product for the Copernicus Emergency Management Service, Publications Office of the European Union, Luxembourg, https://doi.org/10.2760/653891, 2020.

Mazzetti, C., Carton de Wiart, C., Gomes, G., Russo, C., Decremer, D., Ramos, A., Grimaldi, S., Disperati, J., Ziese, M., Schweim, C., Sanchez Garcia, R., Jacobson, T., Salamon, P., and Prudhomme, C.: River discharge and related historical data from the European Flood Awareness System, v5.0, https://doi.org/10.24381/cds.e3458969, 2023.

Molinari, D., De Bruijn, K. M., Castillo-Rodríguez, J. T., Aronica, G. T., and Bouwer, L. M.: Validation of flood risk models: Current practice and possible improvements, International journal of disaster risk reduction, 33, 441–448, 2019.

Paprotny, D., Terefenko, P., and Śledziowski, J.: HANZE v2.1: an improved database of flood impacts in Europe from 1870 to 2020, Earth System Science Data, 16, 5145–5170, 2024.

Paprotny, D., Hart, C. M. P., and Morales-Nápoles, O.: Evolution of flood protection levels and flood vulnerability in Europe since 1950 estimated with vine-copula models, Natural Hazards, 121, 6155–6184, 2025.

Pianosi, F., Sarailidis, G., Styles, K., Oldham, P., Hutchings, S., Lamb, R., and Wagener, T.: Towards global sensitivity analysis of large-scale flood loss models, Natural Hazards and Earth System Sciences, 26, 1727–1743, https://doi.org/10.5194/nhess-26-1727-2026, 2026.

Pigaiani, C. and Batista E Silva, F.: The LUISA Base Map 2018, https://doi.org/10.2760/503006, 2021.

Platt, S., Carpenter, O., Mahdavian, F., and Coburn, A.: Disaster recovery–evidence from 100 natural disasters, International Journal of Disaster Risk Reduction, p. 105764, 2025.

https://doi.org/10.5194/egusphere-2026-2807
Preprint. Discussion started: 1 June 2026

Portalés-Julià, E., Mateo-García, G., Purcell, C., and Gómez-Chova, L.: Global flood extent segmentation in optical satellite images, Scientific reports, 13, 20 316, 2023.

Riedel, L., Röösli, T., Vogt, T., and Bresch, D. N.: Fluvial flood inundation and socio-economic impact model based on open data, Geoscientific Model Development, 17, 5291–5308, 2024.

Risling, A., Lindersson, S., and Brandimarte, L.: A comparison of global flood models using Sentinel-1 and a change detection approach, Natural Hazards, 120, 11 133–11 152, 2024.

Rojas, R., Feyen, L., and Watkiss, P.: Climate change and river floods in the European Union: Socio-economic consequences and the costs and benefits of adaptation, Global Environmental Change, 23, 1737–1751, 2013.

Salamon, P., Mctlormick, N., Reimer, C., Clarke, T., Bauer-Marschallinger, B., Wagner, W., Martinis, S., Chow, C., Böhnke, C., Matgen, P., et al.: The new, systematic global flood monitoring product of the copernicus emergency management service, in: 2021 IEEE international geoscience and remote sensing Symposium IGARSS, pp. 1053–1056, IEEE, 2021.

Scarpellini, L., Ficchì, A., D'Angelo, C., Betterle, A., Salamon, P., and Castelletti, A.: [Dataset] Mapping uncertainty in flood impacts: a multi-dataset assessment and catalog of simulated, remote sensed and reported floods in Europe, https://doi.org/10.5281/zenodo.19666073, 2026.

Schiavina, M., Freire, S., Carioli, A., and MacManus, K.: GHS-POP R2023A - GHS population grid multitemporal (1975-2030), https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE, 2023.

Scussolini, P., Aerts, J. C., Jongman, B., Bouwer, L. M., Winsemius, H. C., de Moel, H., and Ward, P. J.: FLOPROS: an evolving global database of flood protection standards, Natural Hazards and Earth System Sciences, 16, 1049–1061, 2016.

Tehrany, M. S., Pradhan, B., Mansor, S., and Ahmad, N.: Flood susceptibility assessment using GIS-based support vector machine model with different kernel types, Catena, 125, 91–101, 2015.

Toosi, A. S., Doulabian, S., Tousi, E. G., Calbimonte, G. H., and Alaghmand, S.: Large-scale flood hazard assessment under climate change: A case study, Ecological Engineering, 147, 105 765, 2020.

Trigg, M., Birch, C., Neal, J., Bates, P., Smith, A., Sampson, C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., et al.: The credibility challenge for global fluvial flood risk analysis, Environmental Research Letters, 11, 094 014, 2016.

Twele, A., Cao, W., Plank, S., and Martinis, S.: Sentinel-1-based flood mapping: a fully automated processing chain, International Journal of Remote Sensing, 37, 2990–3004, 2016.

Van Der Knijff, J., Younis, J., and De Roo, A.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, 2010.

Vergni, L., Todisco, F., and Di Lena, B.: Evaluation of the similarity between drought indices by correlation analysis and Cohen's Kappa test in a Mediterranean area, Natural Hazards, 108, 2187–2209, 2021.

Wing, O. E., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., and Erickson, T. A.: Validation of a 30 m resolution flood hazard model of the conterminous U nited S tates, Water Resources Research, 53, 7968–7986, 2017.

Wing, O. E., Smith, A. M., Marston, M. L., Porter, J. R., Amodeo, M. F., Sampson, C. C., and Bates, P. D.: Simulating historical flood events at the continental-scale: observational validation of a large-scale hydrodynamic model, Natural Hazards and Earth System Sciences Discussions, 2020, 1–30, 2020.

Zhao, G., Yamazaki, D., Tanaka, Y., Zhou, X., Li, S., Hu, Y., Hirabayashi, Y., Neal, J., and Bates, P.: Developing a levee module for global flood modeling with a reach-level parameterization approach, Water Resources Research, 61, e2024WR039 790, 2025.