

Optimizing Warnings for Hydrologic Ensemble Prediction Systems for Improved Decision-Making

JESÚS CASADO-RODRÍGUEZ^a,^b CORENTIN CARTON DE WIART^b, STEFANIA GRIMALDI^a, ERVIN ZSOTER^b,^c CALUM BAUGH^b, NINA BOSSHARD^c, MICHAELA MIKULÍČKOVÁ^d, ILIAS PECHLIVANIDIS^c, CHRISTEL PRUDHOMME^b, AND PETER SALAMON^a

^a European Commission - Joint Research Centre, Ispra, Italy

^b European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

^c Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

^d Slovak Hydrometeorological Institute, Bratislava, Slovakia

(Manuscript received 24 April 2024, in final form 6 November 2024, accepted 27 February 2025)

ABSTRACT: Flood early warning systems (FEWSs) rely on hydrological simulations driven by numerical weather prediction (NWP) models, both of which are inherently uncertain. Ensemble prediction systems (EPSs) address these uncertainties by generating multiple future scenarios. Discharge forecasts from EPS are converted into flood warnings by applying a set of criteria whose definition is crucial for the skill of the system. While meteorological and hydrological models are under continuous development, these warning criteria lack continuous evaluation. In this paper, we use discharge simulations from the European Flood Awareness System (EFAS) to assess the skill of four NWPs—probabilistic and deterministic—and explore methods to create a grand ensemble that enhances overall skill. We examine the effects of the current warning criteria—probability threshold and persistence—optimize their values, and evaluate their effectiveness in smaller catchments. Our results indicate that probabilistic NWPs outperform deterministic models in flood warning skill and demonstrate that removing the persistence criterion from EPS enhances skill. Optimized warning criteria for a single probabilistic NWP boost EFAS skill by 6.6% in terms of *f* score, and a grand ensemble can further increase it by 4.1%. We identify two effective grand ensemble methods—member based and skill based—and discuss their advantages and drawbacks. The improved criteria demonstrate comparable skill in catchments half the size of those currently used. In conclusion, our study presents a methodology for evaluating and refining the skill of multimodal FEWS, stressing the critical role of tailoring the flood warning criteria to the end users.

SIGNIFICANCE STATEMENT: This study analyzes methodologies to improve the skill of the notifications issued by a state-of-the-art flood warning system. We have analyzed the skill of four meteorological models and the effects of several warning criteria. We have identified an optimal blend of meteorological models and warning criteria that boosts the system skill up to 10%. The methods developed in this research are adaptable to other flood warning systems, and the refined criteria have been implemented operationally.

KEYWORDS: Ensembles; Forecast verification/skill; Operational forecasting; Hydrologic models; Flood events

1. Introduction

Floods are the natural disasters that cause the biggest economic losses in Europe (WMO 2021). In the European Union, flood damages amounted to a total cost of approximately EUR 280 billion in the period 1980–2022, representing 43% of all meteorological hazards [European Environment Agency (EEA) 2023]. The historical data show an increasing tendency in flood damages over Europe, and climate projections indicate an exacerbation of this trend with increasing frequency and severity of extreme events (EEA 2023; Paprotny et al. 2018; Alfieri et al. 2017; Caretta et al. 2023). In this scenario, flood early warning systems (FEWSs) emerge as one of the

multiple adaptation measures to be taken. The magnitude of the flood that hit Germany and Belgium in July 2021, the costliest in the last 40 years, was a reminder of the importance of FEWS and their need for continuous improvement of forecasts and decision-making (EEA 2023).

FEWSs rely on meteorological forecasts generated by numerical weather prediction (NWP) models (De Roo et al. 2003; Gouweleeuw et al. 2005). Over the past 30 years, ensemble prediction systems (EPSs) have been developed within the realm of NWP models to address uncertainties in the initial conditions. EPSs provide equiprobable forecasts at any given time and location, enabling the quantification and dissemination of probabilistic forecasts (Cloke and Pappenberger 2009). In comparison with deterministic models, EPSs have demonstrated enhanced consistency across consecutive forecasts and improved precipitation forecasts at longer lead times (Buizza 2008; Molteni et al. 1996).

The success of EPS in meteorology led to the adoption of this paradigm in flood forecasting (De Roo et al. 2003; Bartholmes and Todini 2005; Pappenberger et al. 2005;

^a Denotes content that is immediately available upon publication as open access.

Corresponding author: Jesús Casado-Rodríguez, jesus.casado-rodriguez@ec.europa.eu

DOI: 10.1175/JHM-D-24-0054.1

© 2025 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by Swedish Meteorological & Hydrological Institute | Unauthenticated | Downloaded 06/18/25 07:15 AM UTC

Roulin and Vannitsem 2005; Diomedea et al. 2008; Yang et al. 2020). Hydrological EPSs (HEPSs) propagate the uncertainties in the meteorological forecasts by forcing hydrological models with the members of a meteorological ensemble. Although this approach is the most common, it only considers the uncertainty in the initial meteorological condition but does not include other sources of uncertainty, such as model structure or model parameterization (Cloke and Pappenberger 2009).

Issuing weather-related warnings, such as those for flooding, is a delicate decision as it might result in significant costs—due to either unpreparedness if an event is missed or the mobilization of resources in the case of false alarms. Weather forecasting systems often issue a high number of false alarms, despite the potential risk of eroding trust in the system, known as the “cry wolf” effect (Bouttier and Marchal 2024; LeClerc and Joslyn 2015). This latter research showed that an increase in false alarms diminishes the credibility of the system, leading end users to disregard precautionary measures. However, they also demonstrated that incorporating uncertainty estimates in warning messages can assist users in making informed decisions. Therefore, HEPs have the potential to enhance both the skill and trustworthiness of flood warning systems, if particular attention is given to the balance between false alarms and missed events.

The major flood events in the Oder River in 1997 and the Elbe River in 2002 triggered the development of the European Flood Awareness System (EFAS), a continental-scale FEWS that has been operational since 2012 as part of the Copernicus Emergency Management Service of the European Commission (Bartholmes et al. 2009; Thielen et al. 2009a). EFAS is an operational system that monitors and forecasts floods in an extended European domain. Its primary objective is to enable proactive measures in anticipation of significant flood events. EFAS is a complementary system to national or regional counterparts, with a specific focus on transboundary rivers and medium-range forecasts.

With the release of version 4.0 in October 2020, EFAS has been generating forecasts twice a day with 6-h resolution and lead times extending up to 10 days. It couples four meteorological inputs of diverse nature—deterministic and probabilistic—with the distributed, process-based hydrological model LISFLOOD-OS (De Roo et al. 2000; Burek et al. 2013). The simulation produces a set of water fluxes and state variables, from which river discharge is utilized to issue flood warnings. EFAS employs a threshold exceedance approach for issuing warnings, where continuous discharge time series are converted into binary series of exceedance or nonexceedance over a discharge threshold (Ramos et al. 2007; Thielen et al. 2009a; Bartholmes et al. 2009). These works developed a warning system based on two criteria: the probability threshold, which defines when there is sufficient certainty about the occurrence of an event in a particular forecast, and the persistence of that signal across consecutive forecasts.

The criteria applied in EFAS to issue flood warnings have changed over the past years driven by the experience of the forecasters and feedback from users. Previous studies have examined EFAS hydrological performance (Alfieri et al. 2014), its warning skill (Bartholmes et al. 2009; Thielen et al.

2009b), and the response of end users (Demeritt et al. 2013). However, the continuous advances in NWP performance (Bauer et al. 2015), as well as in the resolution and representation of hydrological processes in EFAS [Joint Research Centre (JRC) 2020, 2023], require a new evaluation of the warning criteria to harness the potential of the system.

The objective of this analysis is to assess the skill of a continental flood forecasting system like EFAS and explore new warning criteria that can maximize that skill. In particular, the first goal is to determine the value of combining several NWPs of different natures in a single system, specifically seeking the appropriate combination method to build a grand ensemble. The second goal is to optimize the two warning criteria, analyzing the interplay between them and reconsidering whether the persistence criterion is meaningful in a purely probabilistic approach. The third and final goal is to analyze the relationship between warning skill and catchment area to discern if the newer, higher-resolution models are skillful in smaller catchments.

To address our research questions, we conducted an analysis of EFASv4 discharge simulations for the period 2020–23. Section 2 enumerates the dataset used in the analysis. Section 3 explains the two experiments that analyze the skill of individual NWPs and different combinations of those models, the criteria optimization process, and the skill metrics we targeted. Section 4 presents first the skill of individual NWP, followed by that of the combination methods, and ends with an exploration of the relationship between warning skill and catchment area. Section 5 discusses the importance of the spatiotemporal framework and target metric in skill assessments, the trade-offs between the warning criteria, the benefits of combining NWP models, and the challenges in selecting a combination method.

2. Data

The data used in this analysis are taken from the operational setup of EFASv4 (JRC 2020). EFAS generates a hydrological forecast twice a day (0000 and 1200 UTC) forced by the four meteorological forecasts outlined in Table 1. EFAS integrates forecasts from probabilistic models (COS and ENS) and deterministic models (DWD and HRES) to issue flood warnings.

The study period spans from the release of EFASv4 in October 2020 until June 2023. We compare EFAS forecasted discharge with the EFAS long-term historical simulation, referred to here as observed discharge. Using the historical simulation as ground truth addresses challenges posed by limited hydrological observations (e.g., gaps in the observed time series and difference in recorded period) and removes model representation and calibration errors from the analysis (Pappenberger et al. 2015; Cloke et al. 2017).

In contrast to the original EFAS skill assessment (Bartholmes et al. 2009), this work focuses solely on real gauging stations rather than all river cells in the model, which limits the spatial correlation. From the approximately 4000 gauging stations in the EFAS database, we filtered stations based on two conditions: The contributing area must exceed 500 km², and the hydrological performance in the historical simulation must not fall below 0.50 in terms of the modified Kling–Gupta efficiency

TABLE 1. Characteristics of the meteorological forecasts used in EFASv4.

Model	Provider	Acronym	Max lead time (days)	No. of ensembles	Spatial resolution (km)
COSMO-LEPS		COS	5.5	20	~7
ICON-EU/ICON	DWD	DWD	7	1	~6.5–13
HRES	ECMWF	HRES	10	1	~9
ENS	ECMWF	ENS	10	51	~18

(KGE) (Kling et al. 2012; Gupta et al. 2009; Knoben et al. 2019). The lower bound on catchment was set due to the relatively coarse spatial resolution of NWP and EFAS temporal resolution (6 h), which limits hydrological performance in small catchments. The condition based on hydrological performance ensures that the hypothesis of using model simulations as proxies for observations is valid. Ultimately, 1979 gauging stations were included in the analysis, accounting for a total of 1683 events. As explained in section 3d, the optimization of the flood warning criteria was conducted on a subset of 1239 stations with a contributing area larger than 2000 km² (874 events), while the entire set was used to analyze the influence of catchment area on warning skill. Appendix A illustrates the distribution of the selected stations and the number of observed events.

3. Methods

The skill assessment comprises four steps, as depicted schematically in Fig. 1: 1) transforming both observed and forecasted

discharge time series into corresponding time series of probability of exceeding a specific return period (Fig. 1a); 2) building an exceedance probability matrix that consolidates the overlapping forecasts, followed by identifying predicted events through the application of the warning criteria (Fig. 1b); 3) creating the contingency table (Fig. 1c); and 4) evaluating the skill of the system across all possible combinations of the warning criteria (Fig. 1d).

a. Predicting flood events

The warning process starts by converting the continuous time series of discharge into time series of exceedance or non-exceedance over a predefined threshold. As forecasts are issued more frequently than the forecast horizon, they overlap. To manage this overlap, the ensemble forecasts are restructured into an exceedance probability matrix, aligning each time step with increasing lead times, as illustrated in Fig. 1b. Within this matrix, an individual forecast is represented by a

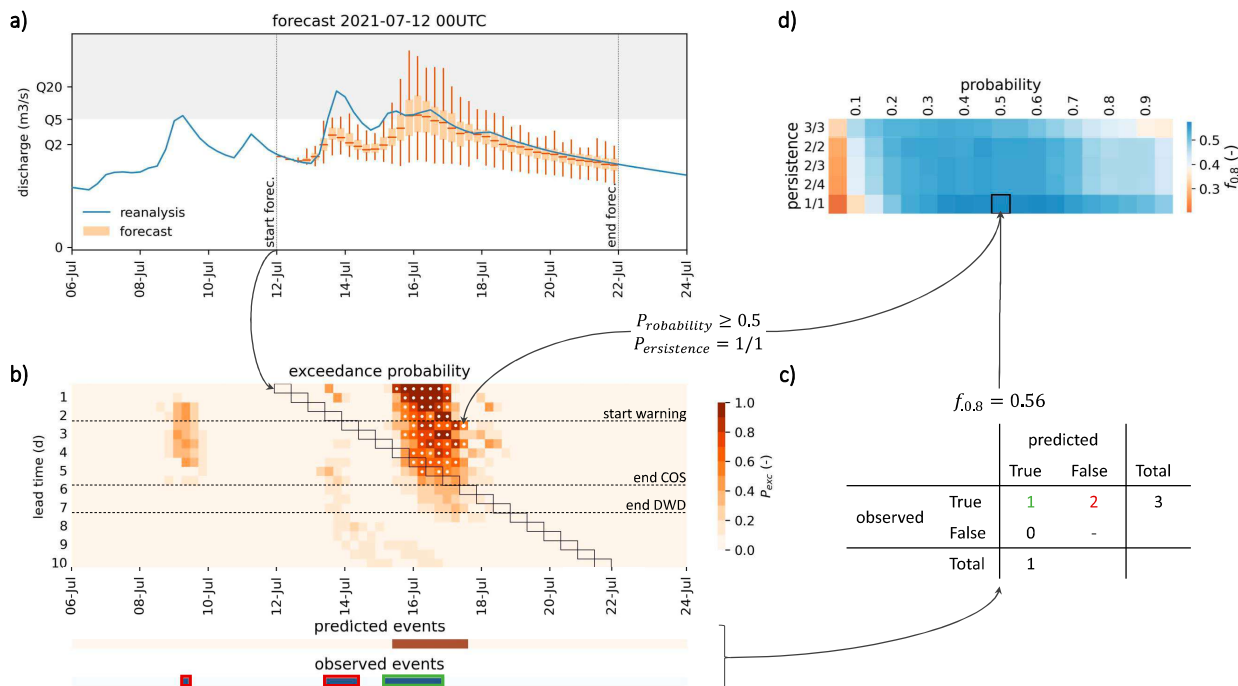


FIG. 1. Overview of the skill assessment framework. (a) Hydrographs comparing observed and forecasted discharge with the discharge associated with the 5-yr return period Q_5 , indicated by gray shading. (b) Matrix of total exceedance probability [white dots identify cells that meet the warning criteria, whereas black rectangles delineate the location of the forecast hydrograph in (a)], timeline of predicted events for that specific criteria, and timeline of observed events where hits are highlighted in green and misses in orange. (c) Confusion matrix based on the previous warning criteria. (d) Evaluation of the f score across all combinations of probability threshold and persistence—the black square highlights the skill score corresponding to the warning criteria in (b).

diagonal (black rectangles). The uppermost rows represent shorter lead times, where the probability of accurate prediction generally increases. For instance, in the example in Fig. 1, the flood event occurring on 16 July was predicted with 50% probability 5 days in advance, increasing close to certainty within 2 days of the event.

The derivation of the binary time series of observed events is straightforward by applying the threshold to the observed discharge. Conversely, deriving the binary time series of predicted events involves applying the warning criteria to the matrix of probability of exceedance. The probability threshold sets the minimum exceedance probability at which a forecasted event is highly certain. Additionally, the persistence criterion removes false alarms caused by the erratic behavior of some NWP (Bartholmes et al. 2009). This criterion replicates the behavior of a forecaster, who would wait to send the warning until consecutive forecasts predict the event. Figure 1b illustrates the application of these two warning criteria on the matrix of exceedance probability. For every time step (columns in the probability matrix), the white dots identify the lead times that meet the criteria. A flood event is predicted at a particular time step if any of the lead times meet the criteria.

EFAS CURRENT WARNING CRITERIA

Flood warnings in EFAS are based on discharge, and the critical threshold is the discharge associated with the 5-yr return period Q_5 , computed by fitting a Gumbel distribution to the annual maxima in the EFASv4 long-term discharge simulation (1990–2022). While alternative alert systems might rely on disparate variables, such as precipitation (Bouttier and Marchal 2024) or river stage (Nevo et al. 2022), and employ different thresholds for defining events, the underlying methodology remains similar.

To issue a flood warning, five criteria must be met. EFAS warnings are exclusively issued to locations with a minimum catchment area of 2000 km² and for lead times longer than 48 h. The first limit is based on the inferior hydrological performance in smaller catchments (Alfieri et al. 2014), while the second limit is a consequence of EFAS addressing early awareness and serving as a supplemental system to national or regional services. A warning is issued if at least one deterministic and one probabilistic NWPs predict a probability of exceeding Q_5 larger than 30% over three consecutive forecasts (hereafter referred to as 3/3 persistence). We denote this approach of combining the NWP models as 1 deterministic + 1 probabilistic (1D + 1P). As each model is evaluated independently, the current procedure does not use a total probability approach.

b. Assessing warning skill

1) CONTINGENCY TABLE

Assessing the skill of a warning system is a binary classification task that compares the time series of predicted and observed events (Fig. 1b). The contingency table (Table 2) summarizes the four possible outcomes. We consider a hit any observed event in which at least one time step was correctly predicted. Then, the number of misses and false alarms is

TABLE 2. Contingency table in a binary classification. The term E_{obs} is the sum of observed events, and E_{pred} is the sum of predicted events.

		Observed		E_{pred}
		True	False	
Forecasted	True	Hit	False alarm	
	False	Miss	True negative	
		E_{obs}		

calculated as the difference between the observed E_{obs} or predicted E_{pred} events and the hits, respectively:

$$\text{misses} = E_{\text{obs}} - \text{hits}, \quad (1)$$

$$\text{false alarms} = E_{\text{pred}} - \text{hits}. \quad (2)$$

2) SKILL METRICS

Flood events are rare, resulting in a number of true negatives that is orders of magnitude larger than any other term in the contingency table. To avoid overestimating skill, we must omit the true negatives in the selection of the target skill metric. Metrics such as the odds ratio and the Hanssen–Kuipers (HK) score (Hanssen and Kuipers 1965), which include true negatives, are therefore unsuitable and were excluded. Instead, we selected three metrics specific for imbalanced classification.

Recall is the proportion of observed events that are correctly predicted [Eq. (3)]. Bartholmes et al. (2009) proved that recall is equal to HK for highly imbalanced classifications. Precision is the proportion of the predicted events that are correct [Eq. (4)]. There is a well-known trade-off between recall and precision. Strict criteria result in fewer, more certain warnings, leading to high precision but low recall due to many missed events. Conversely, relaxed criteria increase the number of warnings, improving recall at the cost of precision due to a rise in false alarms. As a compromise, the f_β score is a metric that balances recall and precision [Eq. (5)]. In its most common version ($\beta = 1$), it is the harmonic mean of recall and precision, granting equal importance to both terms. However, higher importance can be given to precision ($\beta < 1$), limiting the amount of false alarms, or to recall ($\beta > 1$), reducing the number of misses. In our study, after suggestions from forecasters analyzing EFAS on a daily basis, we selected $f_{0.8}$ as the target skill metric. This metric balances precision and recall giving a slightly higher value to the former. The idea behind it is to limit the number of false alarms, which would jeopardize the trust of the recipients of the warnings. Recall, precision, and the f_β score range from 0 to 1, with 1 being their optimal value.

As a final validation metric, we used bias, which represents the proportion between the number of predicted and observed events [Eq. (6)]. Its optimal value is 1, which means that the number of warnings is equal to the number of observed events, and it ranges from 0 to infinity. All four metrics

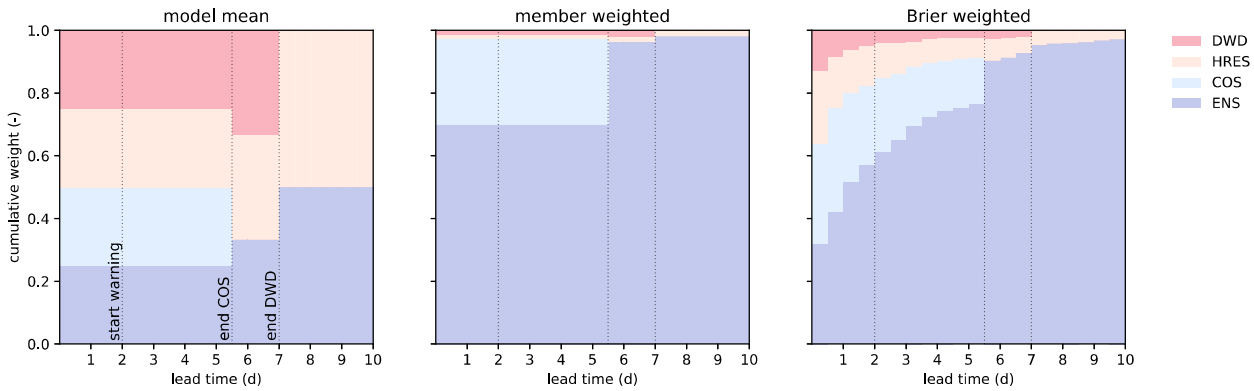


FIG. 2. Weights applied to compute the total probability. Every plot represents a different method, with colors representing NWP (blue for probabilistic models and red for deterministic models).

presented here can be plotted in the Roebber diagram (Roebber 2009) (appendix C):

$$\text{recall} = \frac{\text{hits}}{E_{\text{obs}}} = \frac{\text{hits}}{\text{hits} + \text{misses}}, \quad (3)$$

$$\text{precision} = \frac{\text{hits}}{E_{\text{pred}}} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}, \quad (4)$$

$$f_{\beta} = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \quad (5)$$

$$\text{bias} = \frac{E_{\text{pred}}}{E_{\text{obs}}} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}} = \frac{\text{recall}}{\text{precision}}. \quad (6)$$

c. Combination of NWP

One of the strengths of EFAS is the concurrent use of four meteorological forecasts, both deterministic and probabilistic. Using multiple NWPs serves two purposes: to allow national services to compare their forecasts with the corresponding EFAS simulations and to leverage the strengths of different NWPs. In relation to this second purpose, a primary objective of this study is to develop an effective method for integrating deterministic and probabilistic forecasts.

In practical terms, EFAS forecasts produce four distinct probability matrices—one for each NWP—rather than a single matrix as depicted in Fig. 1b. The challenge lies in devising a methodology to combine these four matrices into a total exceedance probability matrix that enhances the skill of the system (Wetterhall et al. 2013). We compare four combination methods: three approaches for calculating the total exceedance probability and the method currently used in EFAS (section 3a). The difference between the three total probability approaches is the way in which they assign weights to each NWP and time step (Fig. 2).

The model mean (MM) attributes equal weights to each model, regardless of whether it is deterministic or probabilistic. The changes in the weight distribution occur when the maximum lead time of a model is exceeded. In this approach, the single member of a deterministic model is considered

significantly more important than any individual member of a probabilistic model.

Instead, in the member-weighted (MW) combination, every model run is assigned an equal weight. This approach heavily relies on the skill of the largest ensemble, which is not necessarily the most skillful. ENS, which has a significantly larger number of members, prevails over any other NWP.

In the Brier-weighted (BW) combination, weights are based on the probabilistic skill of the NWPs. The Brier score—a probabilistic error metric that ranges from 0 to infinity, with 0 being the optimal value—is used as a measure of probabilistic skill (Brier 1950):

$$\text{BS} = \frac{1}{T} \sum_{t=1}^T (P_{\text{obs},t} - P_{\text{pred},t})^2, \quad (7)$$

where BS is the Brier score of a station, T is the number of time steps, $P_{\text{obs},t}$ is the observed probability of exceedance (either 0 or 1), and $P_{\text{pred},t}$ is the predicted probability of exceedance. We discovered that the results of the BW approach are highly sensitive to how Brier scores are converted into weights. The conversion function must invert BS—as better models, with smaller BS, should receive higher weight—and should be exponential to amplify the differences among NWP. Inverse distance weighting fulfills both conditions [Eq. (8)]. Values of the exponent p from 1 to 9 were tested, and 7 was found to be the optimal exponent:

$$w_{\text{nwp},t} = \frac{\text{BS}_{\text{nwp},t}^{-p}}{\sum_{i=1}^4 \text{BS}_{i,t}^{-p}}. \quad (8)$$

As shown in Fig. 2, the distribution of weights in BW changes with lead time. At very short lead times, the deterministic models demonstrate skill (Fig. B1, in appendix B), accounting for approximately one-third of the total weight. As lead time increases, the probabilistic models—particularly ENS—assume the majority of the weight.

d. Optimizing warning criteria

Issuing a warning involves five criteria: minimum catchment area, minimum lead time, combination of NWP, probability

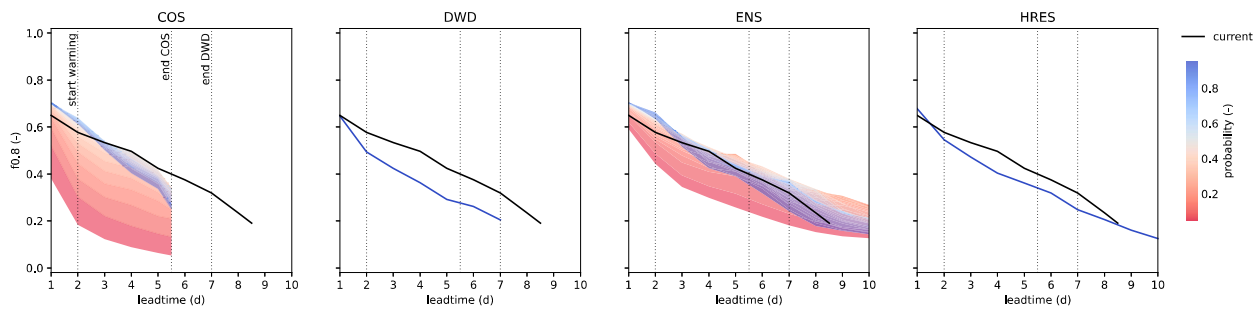


FIG. 3. Evolution of the warning skill across daily lead times and probability threshold for each NWP under a scenario without persistence. Each panel represents a different NWP, with shades of color indicating the spread of model skill over the probability thresholds. The black line serves as a benchmark, indicating the skill of the current operational criteria.

threshold, and persistence. We analyzed each of these criteria individually to understand its effects and then searched for the set of criteria that optimizes the system skill.

We have conducted two main experiments to explore the benefits of combining NWP. The first experiment analyzes each NWP individually with two objectives: identifying the strengths and weaknesses of each NWP and understanding how the warning criteria affect differently probabilistic and deterministic NWPs. The second experiment compares the four combinations of NWP explained in section 3c. The aim is to identify the most appropriate combination method and compare it against the most skillful NWP to assess whether the grand ensemble adds value to the system. The setup of these two experiments is similar and is explained in the following paragraphs.

For the majority of the analyses, we used only stations that met the current minimum catchment area of 2000 km². With this subset of stations, we explored the evolution of skill with lead time, probability, and persistence. Lead times were grouped in two ways. First, we grouped them daily to analyze the evolution of skill with lead time. Second, for the optimization of the warning criteria, we defined four lead time ranges: 1) shorter than 2 days, a range not covered by EFAS warnings; 2) 2–5.5 days, when all NWP are available; 3) 5.5–7 days, when three NWPs are available; and 4) 7–10 days, when only the ECMWF forecasts are available. We tested probability threshold values ranging from 5% to 95% and five persistence values: 1/1 (no persistence), 2/4 (two out of the last four forecasts), 2/3, 2/2, and 3/3 (current criterion).

The optimization of the probability threshold and persistence criteria assessed the warning skill for all possible combinations of these two criteria, identifying the combination that maximizes $f_{0.8}$. This process was repeated for the two experiments—individual NWPs and combinations of NWPs—and the four lead time ranges. To ensure a robust selection of criteria, we applied a 10-fold cross validation: 20% of the stations were randomly left aside as a test set, while the remaining 80% were randomly subdivided into 10 folds. We computed the skill for every combination of nine folds and then averaged over folds. The optimal warning criteria were derived from this average-over-fold skill matrix, as shown in Fig. 1d. Subsequently, the selected criteria were evaluated on the test set to prevent overfitting.

Last, we explored the impact of catchment area on warning skill using the entire set of stations. We calculated the warning skill for increasing values of the minimum catchment area, ranging from 500 to 300 000 km², using the previously optimized criteria.

4. Results

a. Analysis of individual NWPs

1) WARNING SKILL OF INDIVIDUAL NWPs

Figure 3 exhibits the evolution of the warning skill with lead time and probability threshold for a scenario with no persistence. Color shades represent the probability threshold, and the black line represents the skill of the current warning criteria, as a benchmark. The skill of the two deterministic models (DWD and HRES) is represented as a line because it is not affected by the probability threshold.

Overall, the probabilistic models show higher skill than the deterministic ones. Under certain conditions, they also outperform the current warning criteria, indicating potential for enhancing EFAS skill. Deterministic models demonstrate skill comparable to the current criteria solely for the first-day forecast. Warning skill degrades with lead time, a trend consistent across all cases, though the rate of degradation varies with the probability threshold.

Given the degradation of skill with lead time, the selection of the optimal criteria will prioritize the lead time range from 2 to 5.5 days. This approach allows us to mitigate the substantial uncertainty associated with longer lead times and focus on the lead times at which the majority of flood warnings are issued.

2) IMPACTS OF PERSISTENCE ON THE WARNING SKILL OF INDIVIDUAL NWPs

Figure 4 illustrates the evolution of skill with probability threshold and persistence for a fixed lead time range (2–5.5 days). Blue lines represent persistence, and the black cross indicates the current warning criteria, as benchmark. Only three persistence values are shown for clarity. The lines representing deterministic models (DWD and HRES) are horizontal as their skill is not affected by the probability threshold.

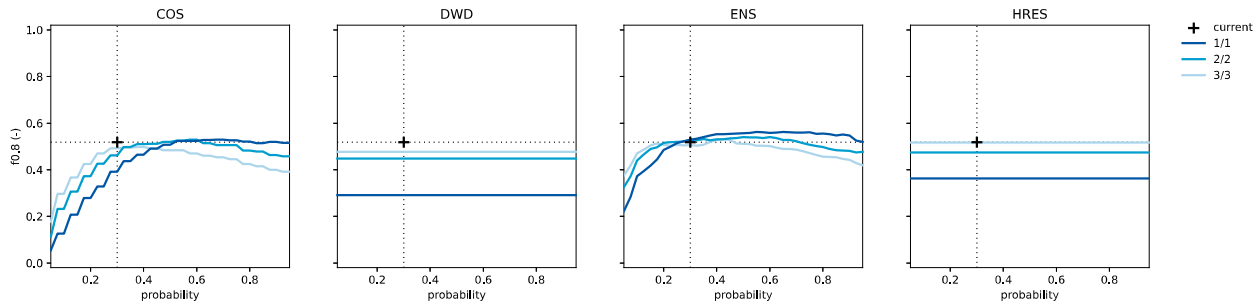


FIG. 4. Evolution of the warning skill with probability threshold and persistence for each NWP at a fixed lead time range (from 2 to 5.5 days). Each panel corresponds to a different NWP, with lines indicating three persistence values. As a benchmark, the black cross represents the current criteria.

Three out of four NWPs replicate or exceed the skill of the current warning criteria at certain combinations of persistence and probability, while only DWD performs below the benchmark. The skill of the probabilistic models shows a trade-off between probability threshold and persistence, as both criteria aim to reduce false positives. Consequently, stricter persistence requires lower probability thresholds to optimize skill and vice versa. Nonetheless, the highest skill is consistently obtained when persistence is removed (1/1). In this scenario, there is a wide range of probability thresholds (from 40% to 90%) that perform similarly well.

Persistence does have a positive impact on the skill of deterministic models. For the most stringent persistence (3/3), the skill of the deterministic HRES matches the benchmark.

3) OPTIMAL WARNING CRITERIA FOR INDIVIDUAL NWPs

The outcome of optimizing the warning criteria for individual NWPs is summarized in Table 3. The results are organized by lead time ranges; for each range, the table presents the

optimized criteria and the skill metrics for the available NWPs. As a benchmark, the first row in each lead time range shows the skill of the current operational criteria. The Roebber diagram in appendix C summarizes visually these results.

The current EFAS setup (1D + 1P) is negatively biased, i.e., prioritizes precision over recall. Warnings are issued with a high level of certainty, leading to a limited number of false alarms, but many events are missed. The current criteria are not optimal, as they are consistently outperformed in terms of $f_{0.8}$ at all lead times by individual NWPs, suggesting that the current combination of all NWPs hinders the skill of the best-performing NWP.

The analysis of the results is similar for any of the four lead time ranges. The probabilistic NWPs (COS and ENS) outperform in terms of $f_{0.8}$ both the benchmark and the deterministic counterparts. Even though both probabilistic NWPs show similar skill at the shortest lead time, ENS stands out as the best model at medium-range forecasts in terms of $f_{0.8}$. The skill of HRES, despite being deterministic, is comparable to the benchmark and even outperforms it at the longest lead time. Overall, the optimization of individual NWPs was able

TABLE 3. Summary of the optimization of the warning criteria for individual NWP and four lead time ranges. The initial row serves as a benchmark, indicating the skill of the current warning criteria. In each lead time range, bold fonts highlight the highest skill for a specific metric.

Lead time	Model	Probability	Persistence	Recall	Precision	Bias	$f_{0.8}$
2 days	1D + 1P	0.300	3/3	0.522	0.735	0.710	0.634
	COS	0.875	1/1	0.673	0.718	0.937	0.700
	DWD	—	2/2	0.585	0.583	1.003	0.583
	ENS	0.800	1/1	0.681	0.711	0.958	0.699
	HRES	—	2/2	0.580	0.671	0.864	0.632
2–5.5 days	1D + 1P	0.300	3/3	0.413	0.618	0.668	0.518
	COS	0.725	1/1	0.387	0.687	0.560	0.518
	DWD	—	3/3	0.420	0.522	0.805	0.477
	ENS	0.650	1/1	0.416	0.725	0.574	0.562
	HRES	—	3/3	0.416	0.612	0.680	0.517
5.5–7 days	1D + 1P	0.300	3/3	0.215	0.612	0.351	0.356
	DWD	—	2/2	0.291	0.358	0.813	0.328
	ENS	0.450	1/1	0.264	0.605	0.436	0.402
	HRES	—	2/2	0.273	0.429	0.636	0.351
	1D + 1P	0.300	3/3	0.120	0.625	0.192	0.237
7–10 days	ENS	0.175	2/2	0.287	0.396	0.725	0.345
	HRES	—	2/2	0.244	0.336	0.726	0.293

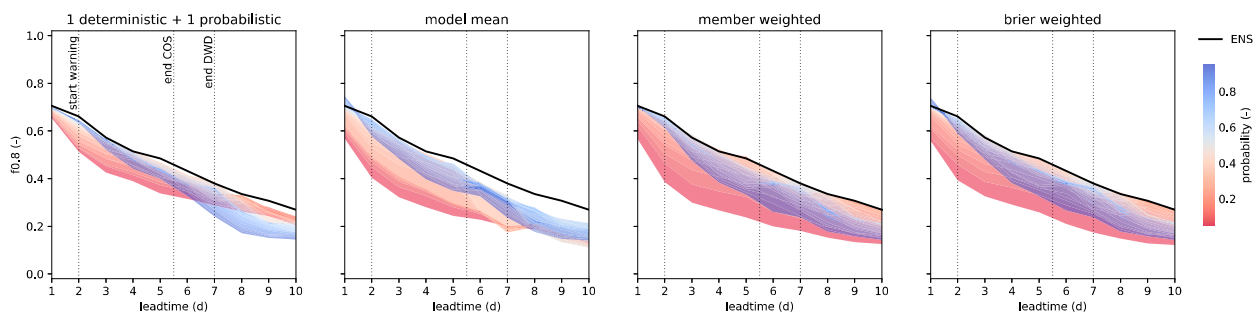


FIG. 5. Evolution of the warning skill across daily lead times and probability threshold for each combination method under a scenario without persistence. Each panel represents a different combination of NWP, with shades of color indicating the spread of model skill over the probability thresholds. The black solid line serves as a benchmark, indicating the skill of the best NWP (ENS with optimized criteria).

to increase the warning skill by up to 6.6% at the shortest and most accurate lead time range (difference in $f_{0.8}$ between the best-performing model and the benchmark).

The optimization removed the persistence criterion for the probabilistic NWPs, except for ENS at the longest lead time range. Its role—removing uncertain events—is now played solely by the probability threshold. The optimal value of this probability threshold is very high at short lead times and decreases with increasing lead time, being larger than the current value up to a lead time of 7 days. On the contrary, deterministic models require persistence to enhance their skill, although there is not a clear pattern between the optimal persistence and lead time.

b. Analysis of combined NWPs

1) WARNING SKILL OF COMBINED NWPs

Figure 5 replicates the analysis of the warning skill previously done for the individual NWP models, but for the combination methods. It exhibits the evolution of the warning skill with lead time and probability threshold for a scenario without persistence. The benchmark is ENS, the NWP that proved the highest skill.

The results indicate that ENS is responsible for the majority of the skill in the grand ensembles, since only marginal gains are possible with specific combinations at certain lead times. Only two combination methods (MW and BW) are able to replicate the skill of the benchmark across all lead times.

Although MM at lead time 1 achieves the highest marginal gain, this combination method is not suitable as its skill is poorer than the benchmark from day 2 onward. The current procedure (1D + 1P) is also suboptimal, even though the degradation with lead time is not as severe as in MM.

2) IMPACTS OF PERSISTENCE ON THE WARNING SKILL OF COMBINED NWPs

Figure 6 presents the evolution of skill for the combinations of NWP across probability and persistence at the lead time range from 2 to 5.5 days. Lines represent different persistence values, and the black cross indicates the skill of ENS with its optimal warning criteria.

The results reinforce the idea that the persistence criterion limits the maximum attainable skill in ensemble systems. Across all four combinations, the highest skill is reached when persistence is removed (1/1). In this scenario, all combinations exhibit similar maximum skill. Only MW and BW marginally surpass the benchmark (ENS). These two combinations show optimal skill within a range of probability thresholds spanning 40%–70%.

The current criteria show a notable lack of sensitivity to probability, as indicated by the lightest blue line in the left-hand panel. Even at the lowest probability threshold, the $f_{0.8}$ score remains at 0.5, unlike the scenario without persistence, that exhibits a significant decline in performance when the probability threshold takes very low values.

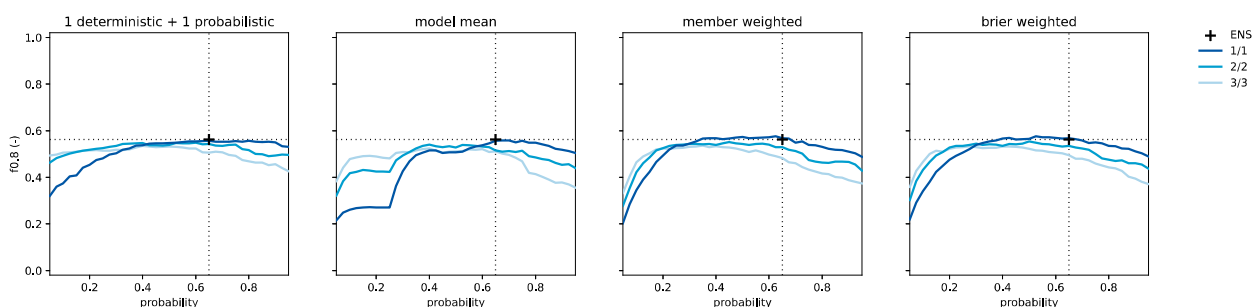


FIG. 6. Evolution of the warning skill with probability threshold and persistence for each combination of NWP at a fixed lead time range (from 2 to 5.5 days). Each panel represents a different combination of NWP, with lines indicating three persistence values. As a benchmark, the black cross represents the skill of ENS with optimized criteria.

TABLE 4. Summary of the optimization of the warning criteria for the combination methods and four lead time ranges. The initial row serves as a benchmark, indicating the skill of ENS, the top-performing NWP. In each lead time range, bold fonts highlight the highest skill for a specific metric.

Lead time	Method	Probability	Persistence	Recall	Precision	Bias	$f_{0.8}$
2 days	ENS	0.800	1/1	0.681	0.711	0.958	0.699
	1D + 1P	0.925	1/1	0.662	0.717	0.923	0.694
	MM	0.775	1/1	0.628	0.838	0.749	0.741
	MW	0.750	1/1	0.673	0.718	0.937	0.700
	BW	0.950	1/1	0.620	0.829	0.748	0.733
2–5.5 days	ENS	0.650	1/1	0.416	0.725	0.574	0.562
	1D + 1P	0.600	1/1	0.455	0.643	0.708	0.554
	MM	0.700	1/1	0.424	0.700	0.606	0.559
	MW	0.500	1/1	0.455	0.692	0.658	0.574
	BW	0.525	1/1	0.451	0.700	0.644	0.576
5.5–7 days	ENS	0.450	1/1	0.264	0.605	0.436	0.402
	1D + 1P	0.425	1/1	0.262	0.599	0.437	0.399
	MM	0.500	1/1	0.276	0.412	0.670	0.345
	MW	0.425	1/1	0.271	0.579	0.468	0.401
	BW	0.450	1/1	0.268	0.586	0.457	0.400
7–10 days	ENS	0.175	2/2	0.287	0.396	0.725	0.345
	1D + 1P	0.250	1/1	0.256	0.414	0.618	0.334
	MM	0.225	2/2	0.252	0.334	0.754	0.296
	MW	0.175	2/2	0.278	0.402	0.692	0.343
	BW	0.300	1/1	0.255	0.445	0.573	0.345

3) OPTIMAL WARNING CRITERIA FOR COMBINED NWPs

Table 4 summarizes the results of the optimization of the warning criteria for the combination methods in an identical way as Table 3 did for the NWP. The results are organized by lead time ranges; for each range, the table presents the optimized criteria and the skill metrics for the four methods. As a benchmark, the first row in each lead time range shows the skill of ENS. A graphical summary of these results can be seen in appendix C.

The benchmark (ENS) is only outperformed in terms of $f_{0.8}$ by some combinations at lead times shorter than 5.5 days. At the shortest lead time range, the inclusion of deterministic NWPs is important as they have notable skill (see Fig. B1). For this reason, the two combination methods that allocate higher weight to deterministic models (MM and BW) show the highest skill. At lead times from 2 to 5.5 days, BW and MW stand out as the best approaches. They have comparable skill and similar optimal criteria due to the fact that they give a similar weight to ENS, as it is both the NWP with the largest number of members and with the highest skill. It is worth noting that the combination currently used in EFAS (1D + 1P) is the only approach that never outperforms the benchmark. Overall, the optimal grand ensemble can achieve an improvement in skill of up to 4.2% compared with ENS (up to 10.7% compared with the current criteria) at the shortest lead time range.

The optimized criteria exhibit a pattern closely aligned with that observed in individual NWPs. For lead times up to 7 days, the persistence criterion is excluded for all combinations, substituted by a probability threshold that adopts notably higher optimal values compared to the current criteria. These values decrease progressively as lead time increases.

4) GEOGRAPHICAL DISTRIBUTION OF WARNING SKILL

The maps in Fig. 7 illustrate the skill metrics associated with the best grand ensemble—BW with no persistence and a probability threshold of 52.5%—for the lead time range of 2–5.5 days.

A pronounced contrast is evident in the maps, attributable to the limited number of events in most of the stations, resulting in skill metrics primarily yielding values of either 1 or 0. In terms of $f_{0.8}$, there is a cluster of underperforming points in central Europe (notably the Alps) and central France. The decomposition into recall and precision reveals that the primary contributor to the low f score is recall, indicating missed events. Conversely, the precision of warnings remains consistently high, with exceptions observed in the Seine and Arno rivers. It is noteworthy the skill in all metrics for the Ebro, Rhine, and Meuse—three rivers that experienced significant flood events during the study period.

By applying the optimized criteria, EFAS would be able to predict with at least 2 days of lead time 46% of the events, and 71% of all the warnings would be correct. For reference, these figures represent an improvement in both metrics over the existing criteria, which achieved 41% recall and 62% precision, respectively.

5) INFLUENCE OF CATCHMENT AREA ON WARNING SKILL

The preceding sections presented results for a set of stations with a catchment area of at least 2000 km², the current limit in EFAS. To investigate the potential relaxation of this criterion, Fig. 8 illustrates the evolution of skill depending on the catchment area limit. It compares the skill of the current criteria with the best grand ensemble—BW with no persistence and a

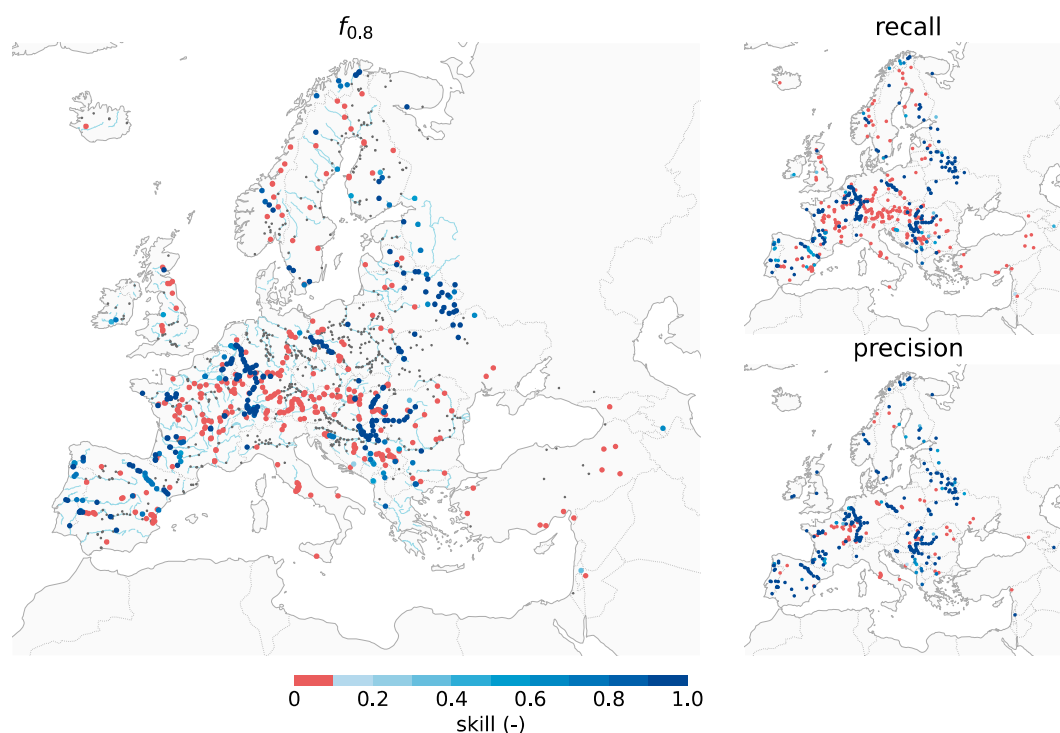


FIG. 7. Skill maps for the optimized warning criteria (BW, 52.5% probability and no persistence) for lead times ranging from 2 to 5.5 days. Each panel represents a metric: the term $f_{0.8}$ in the main panel and its components recall and precision in the smaller panels. The color scale indicates the optimized skill. Gray dots on the main panel denote stations where the f score cannot be computed due to hits, misses, and false alarms being null.

probability threshold of 52.5%—in the lead time range of 2–5.5 days.

The skill of the system improves with catchment area; however, this improvement is not continuous. Recall drops for areas ranging from 30 000 to 70 000 km², leading to a decrease in $f_{0.8}$. This decline is caused by missed events in the Guadiana, Seine, Loire, Rhone, and Danube catchments. As already mentioned in previous sections, there is a notable gap between

recall and precision in both the optimized and current warning criteria, with precision consistently surpassing recall.

The vertical line at 2000 km² illustrates the gain in skill achieved with the optimization, which is sustained across all catchment area values. Reducing the catchment area limit results in minimal loss in $f_{0.8}$, suggesting that this criterion could be relaxed without compromising skill. For instance, the BW combination at 1000 km² exhibits slightly higher skill ($f_{0.8} = 0.531$)

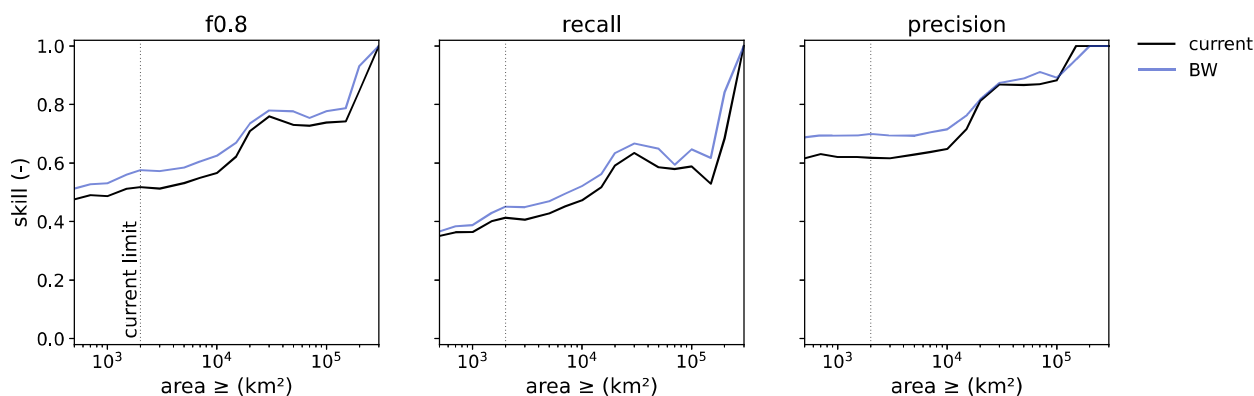


FIG. 8. Evolution of warning skill with catchment area limit at lead times ranging from 2 to 5.5 days. The black line represents the skill of the current warning criteria, and the blue line represents that of the best grand ENS (BW, 52.5% probability and no persistence). The vertical, dotted line indicates the catchment area limit currently applied in EFAS.

than the current criteria at 2000 km² ($f_{0.8} = 0.518$). The relaxation of the catchment area limit impacts the components of the f score differently. While precision remains unaffected, there is some deterioration in recall.

5. Discussion

a. Limitations and moving forward

The definition of the spatial and temporal framework is often a limitation in studies that assess the skill of ensemble flood forecasting systems (Pappenberger et al. 2008b; Cloke and Pappenberger 2009). Due to the rarity of flood events, the statistical robustness of such analysis relies on a limited and often highly correlated sample of events. To address this limitation, long-term and continental or global studies are advocated, such as the one presented here, which spans over a period of two and a half years and extends over a European domain. Nevertheless, our study would benefit from a larger sample of points and flood events. As shown in Fig. 7, there are many stations for which the skill metrics cannot be calculated due to null hits, misses, or false alarms. A larger sample of observed events would reduce the susceptibility of the analysis to this issue. Two potential solutions emerge: extending the study period, which is limited by the availability of the HEPS forecasts, or using an event threshold of higher recurrence. We decided to work with a sample of stations, instead of all the river cells (Bartholmes et al. 2009), to limit the spatial correlation of events.

In the optimization of the warning criteria, we set a catchment area limit of 2000 km², a threshold inherited from previous EFAS setups with lower spatial and temporal resolution. The results of this study demonstrate that the current resolution of both NWP and hydrological models can produce comparably skillful warnings at smaller catchments (Fig. 8), which can significantly increase the number of stations and flood events analyzed (Fig. A2). In line with this, the new EFASv5 (JRC 2023) released in September 2023 increased spatial resolution from 5 km to 1 arc min, which is expected to further improve the skill of the system in smaller catchments. The skill of this new EFAS version will be analyzed as soon as a long enough period or reforecasts are available.

The selection of the target metric is a critical aspect in any skill analysis. Bartholmes et al. (2009) reviewed the desired characteristics of a skill score and selected odds ratio, HK score, and bias as their target metrics. In contrast, we selected the f_β score, a metric commonly used in machine learning, but rarely applied in hydrology. The f_β score has two interesting properties for analyzing a flood warning system. First, it is based on the contingency table but avoids using true negatives, which could overestimate skill. Second, it allows for tailoring the warning criteria to the specific use case, as the β parameter can be adjusted to penalize differently misses or false alarms. For instance, Bouttier and Marchal (2024) used the f_2 score to optimize the probability threshold for high-intensity precipitation warnings in France. The β value of 2 penalizes four times more missed events than false alarms, under the assumption that the cost of being hit by an extreme

event unprepared is larger than the cost of mobilizing resources in a false alarm. In our study, we took the opposite approach and decided to minimize the number of false alarms by using a β value of 0.8. This decision was based on the fact that EFAS complements the national or regional systems by issuing medium-range flood warnings. The aim is to alert authorities, not the general public, only when there is substantial evidence that the event will occur, to avoid excessive warnings that could undermine their trust in the system. Additionally, the event threshold (5-yr return period) corresponds to a recurrence time shorter than the flood protection levels in many European regions, so we assume that missing minor events will not be as costly as Bouttier and Marchal (2024) argue. In a preliminary analysis, we tested β values of 0.8, 1, and 1.25 and found that the changes in the f score were minimal (in the order of 3%), and in none of the cases, persistence was needed. However, the selection of β affects the optimal probability threshold and the resulting bias, i.e., the distribution of errors between misses and false alarms. Such decision should be taken in collaboration with the end users to tailor the warnings to their specific needs (Lavers et al. 2020; Dasgupta et al. 2025).

b. Addressing study objectives

One of the objectives of this study was to evaluate the skill of the different NWPs within EFAS, with a specific focus on comparing deterministic and probabilistic models. Figure B1 demonstrates that probabilistic NWPs outperform deterministic models in terms of Brier skill score across all lead times. Only at short lead times do deterministic NWPs reach similar levels of skill. The outcome is similar when applying the warning criteria and measuring warning skill in terms of $f_{0.8}$ (Table 3). At their optimal warning criteria, the probabilistic NWPs, particularly ENS, outperform deterministic NWPs. The skill of both types of NWP degrades with increasing lead time, casting doubt on their ability to issue flood alerts more than 7 days in advance. The effects of the probability threshold and persistence criteria differ depending on the nature of the NWP. The probability threshold has a substantial effect on the skill of probabilistic models, while it does not impact deterministic models. Both COS and ENS perform better at probability thresholds over 50%, with a wide range of similarly performing values. On the contrary, persistence proves to be beneficial in improving the warning skill of deterministic NWPs, but it limits the skill of probabilistic models. Persistence, also known as poor man's ensemble, is a method of creating an ensemble out of a deterministic forecast (Cloke and Pappenberger 2009). For instance, a persistence of 3/3 is an ensemble of three members with a probability threshold of 100%. Our results show that while persistence removes the inherent erratic behavior of deterministic models and notably improves their skill, it should not be applied to HEPS.

Another objective of this study was to determine the appropriate method to combine several NWP models into a grand ensemble. Careful consideration is necessary in selecting this method to avoid diminishing the skill of the most proficient NWP within the grand ensemble. For instance, the current EFAS warning criteria limit the skill of the system to that of the highest-performing deterministic model (HRES). This

limitation is caused by a combination method that excessively relies on deterministic forecasts, as well as the use of persistence. The challenge lies in finding a combination that enhances the skill of the best-performing NWP, which in the case of EFAS is ENS.

Our findings demonstrate that ENS is a very strong baseline, with only a few combination methods yielding gains to its skill. These findings contrast slightly with Pappenberger et al. (2008a), who showed for a specific flood event in Romania in 2007 that grand ensembles can detect extreme events that a single EPS could entirely miss. This different outcome might be attributed to the broader range of EPS they used compared to EFAS, where COS and ENS are not fully independent. In our study case, the effectiveness of the grand ensemble is limited to the first 5 days of lead time. Overall, the most promising approaches are the member-based (MW) and skill-based (BW) methods. Both methods have similar optimal warning criteria (no persistence and a probability threshold close to 50%), and their success is based on granting most of the weight to ENS. However, the reason why they allocate this weight to ENS differs. In the case of MW, it is coincidental that the most skillful model also possesses the largest number of members. On the contrary, BW grants greater value to ENS because it proved to be a superior model. If a more skillful model with fewer members were to be added to the grand ensemble, BW would be able to extract that skill, while MW would not. Therefore, we argue that skill-based combination methods are preferable.

The last objective was to assess whether the catchment size limit could be relaxed. Although the skill of the system degrades with decreasing catchment size, the optimal warning criteria enable comparable skill to the current setup at smaller catchments.

c. Toward operationalization

The decision on the most appropriate combination method cannot be based solely on skill metrics but also consider other factors such as the operational implementation effort, user comprehensibility, or the impact of adding or removing an NWP. Skill-based approaches, such as BW, are theoretically superior as they allocate weights based on quality rather than quantity. They are also the most robust when it comes to adding or removing NWPs. However, they are more challenging to implement and to explain to end users. To derive the matrix of weights (Fig. 2), a skill-based approach requires an extensive period of reforecasts upfront, posing a substantial computational burden on system implementation. Additionally, practitioners using the forecasts would require additional training to understand that the ensemble members are not equiprobable, which would be more intuitive.

There remains a question on the use of deterministic models in a grand ensemble dominated by a probabilistic NWP. When using a member-based combination, such as MW, deterministic models could be removed from the grand ensemble, as the weight allocated to them is so small that their inclusion barely affects the total probability. Skill-based approaches, instead, allocate larger weights to these models in

the very first lead times, when they have proven skill given their often-higher resolution. The inclusion of deterministic models is a topic for future analysis, as the spatial resolution of deterministic and probabilistic NWPs is nowadays similar, and centers such as ECMWF plan to discontinue their deterministic models in favor of the ensembles.

6. Conclusions

This study describes a procedure to optimize warning thresholds in HEPS. We investigated the ability of deterministic and probabilistic NWP models in providing correct flood warnings, and we explored the effects of the warning criteria, namely, probability threshold and persistence. The final objective of the study is to devise a method to combine several NWPs into a grand ensemble that outperforms the most proficient, individual NWP. Our study case is the European Flood Awareness System (EFASv4), a medium-range continental flood warning system.

The individual analysis of the NWPs demonstrated that probabilistic models, particularly ENS, provide better flood warnings than deterministic models, even for short lead times, where the latter show some skill. We discovered that the persistence criterion is useful to increase the skill of deterministic models, as it creates a pseudoensemble, but is counterproductive in the case of probabilistic NWPs. The criterion that maximizes ensemble skill is the probability threshold, whose optimum shows a certain degree of uncertainty, but always exceeds a value of 50%.

By default, the combination of several NWP into a grand ensemble does not improve the overall skill of the flood warning system. As an example, the current EFAS warning criteria, which uses four meteorological forecasts, is outperformed by only using one NWP (ENS) with the adequate criteria. Two of the approaches tested in this work yielded gains in skill compared with the baseline in medium and short lead times. One of these approaches (member weighted) assumes equiprobability among all the model runs, and the other one (Brier weighted) assigns weights to each NWP based on their probabilistic skill. In our study case, these two approaches yielded similar results because the most proficient NWP is also the one with the largest number of members. Despite that, we argue that the skill-based approach is conceptually more appropriate, gives significance to the inclusion of deterministic NWP in the grand ensemble, and would perform better if the set of NWPs would change. However, other study cases may consider other factors such as implementation costs or explainability in the selection of the combination method.

Finally, we explored the effects of catchment area in the warning skill, with the idea of relaxing the current minimum catchment area limit. The results indicate that skill deteriorates with decreasing catchment area, as other studies report, but the improvements in meteorological and hydrological models over the last decade allow for a reduction in the minimum catchment area. We have proven that, with the optimal criteria, we can halve the catchment size limit, thereby expanding coverage across Europe, and achieving performance comparable to the current system setup. We expect that the

skill at smaller catchments will be positively affected by the release of EFASv5 with a higher spatial resolution.

The results of this study have been integrated into EFAS version 5.2, available since 13 August 2024 (JRC 2024). This version includes new total probability products calculated using the Brier-weighted method and a streamlined warning system with a 50% probability threshold for catchments over 1000 km² and lead times up to 7 days.

Acknowledgments. The authors thank Jutta Thielen-del Pozo (JRC) and two anonymous reviewers for their revision and contribution to the original manuscript.

Data availability statement. The EFASv4 historical and forecast discharge data are available at the Early Warning Data Store of the Copernicus Emergency Management Service (<https://ewds.climate.copernicus.eu>). All the scripts, pre-processed data, and results of the study can be found in the GitHub repository at https://github.com/casadoj/EFAS_skill.

APPENDIX A

Analysis of Observed Events

Figure A1 depicts the spatial distribution of the stations used in the optimization of the warning criteria, where colors represent the amount of observed flood events.

The optimization of the warning criteria is based on 1239 stations and 874 observed events. Throughout the study period, 61% of the stations did not exceed the 5-yr return period, which means that hits and misses cannot exist, and the only term in the contingency table can be false alarms.

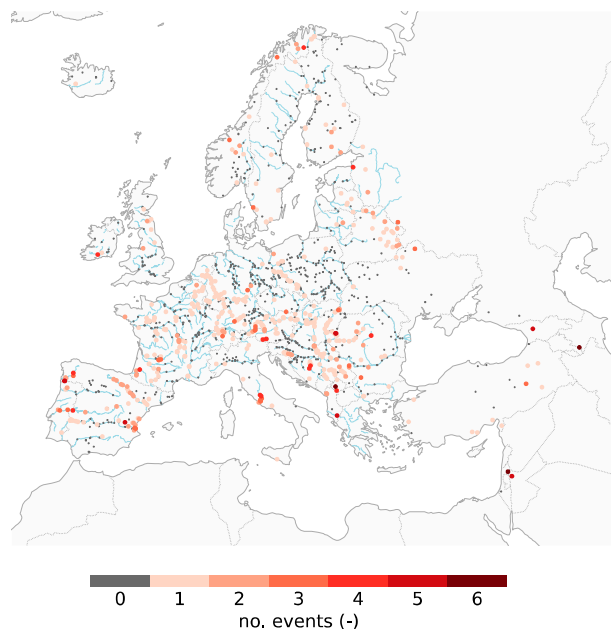


FIG. A1. Geographical distribution of the stations used for optimizing the warning criteria and number of observed flood events during the study period.

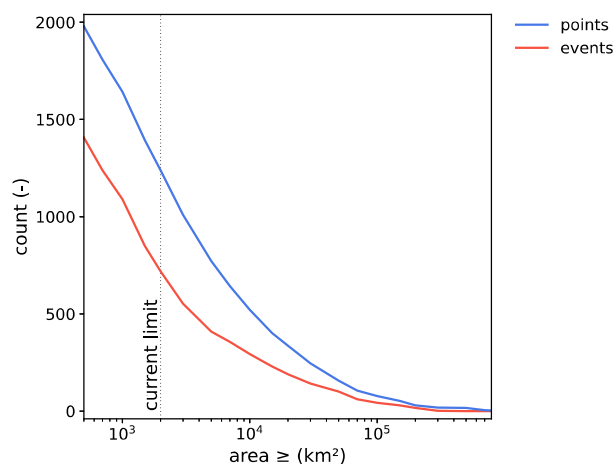


FIG. A2. Number of stations (blue) and observed events (red) by catchment area threshold. The vertical, dotted line represents a catchment area of 2000 km², the current limit in the EFAS warnings.

Consequently, the f score for these stations can be either 0 or null.

The imposition of a minimum contributing area of 2000 km² places a constraint on the number of stations available for optimization, consequently affecting the count of observed events. In the final phase of our analysis, we assessed the implications of this criterion by expanding our consideration to stations with a minimum area of 500 km². Figure A2 represents the relationship between an increasing catchment area threshold and the corresponding number of stations and observed events. By excluding the points with contributing area smaller than 2000 km², we discarded 37% of the stations and 48% of the events. The objective of this decision was to use the current warning criteria as a benchmark.

APPENDIX B

Probabilistic Skill of Individual NWP

In the initial assessment of NWP skill, we calculated the Brier score for each NWP at 12-h lead times. The Brier score, as an error metric, yields values that can be challenging to interpret, given the dependence of error magnitude on the specific variable—particularly in the context of rare events like floods, where the magnitude is inherently low. Instead, we employ the Brier skill score (BSS) to gauge the relative skill of a model in comparison with a reference. Models outperforming the reference receive a positive BSS, while those underperforming register a negative BSS. For the reference, we opted for a dummy model that never predicts an event ($P_{\text{pred}} = 0$) (Legg and Mylne 2004).

Figure B1 demonstrates that probabilistic models exhibit greater skill than deterministic ones. Only within very short lead times (0–1 days) do deterministic models approximate the skill of probabilistic models. This is particularly significant for EFAS warnings, as warnings are issued only at lead times longer than 2 days (indicated by the leftmost

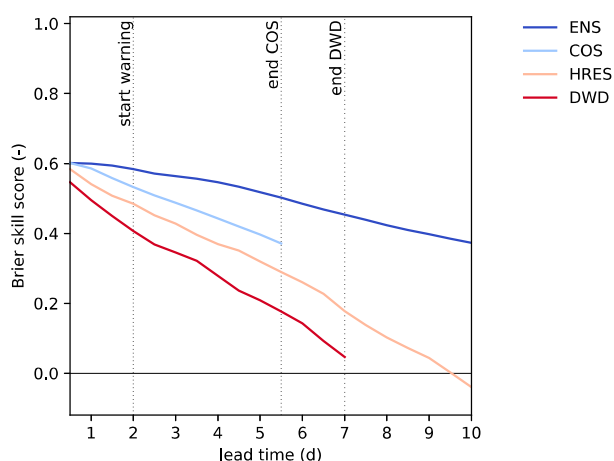


FIG. B1. BSS of the four NWP used in EFAS at 12-h lead times. The reference (BSS = 0) is a model that never predicts an event. Blue lines represent probabilistic models, and red lines represent deterministic ones.

dotted line). As lead time increases, there is a degradation in skill. This decline impacts ENS to a lesser extent than the other models. Both deterministic models display poor skill at their forecast horizon.

APPENDIX C

Roebber Diagram: A Visual Summary of the Results

The Roebber diagram in Fig. C1 offers an alternative perspective on the results of the optimization of the warning criteria. Roebber diagrams condense four skill metrics into a single plot, with an ideal model positioned at the top-right corner (Roebber 2009). Figure C1 shows simultaneously the

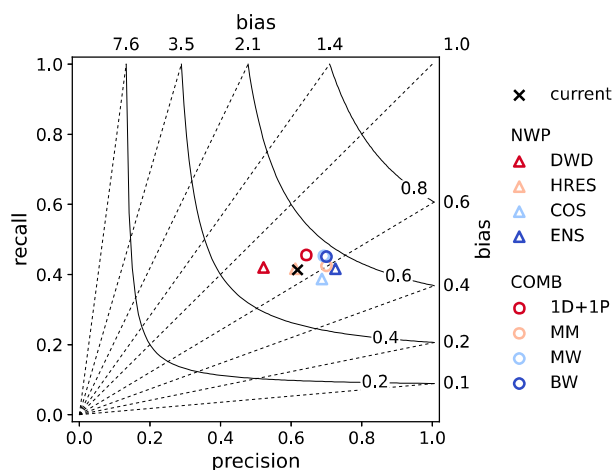


FIG. C1. Roebber diagram of the optimized criteria at the lead time range from 2 to 5.5 days. Four skill metrics are represented: precision and recall in the x and y axes, bias by the dashed lines, and the $f_{0.8}$ by the solid lines. Triangles illustrate individual NWP, and circles illustrate the combination of NWP. For reference, the skill of the current warning criteria is shown as a black cross.

skill of the current warning criteria, the optimization of individual NWP (Table 3), and the optimization of combination methods (Table 4) for the lead time range of 2–5.5 days.

All NWP models and the current criteria exhibit negative bias, indicating that EFAS underpredicts the occurrence of flood events. In the case of the optimizations carried out in this analysis, the selection of the $f_{0.8}$ score as the target metric in the optimization forces this bias, as it prioritizes precision over recall. It is remarkable that the improved $f_{0.8}$ of the probabilistic models is a result of higher precision, while recall values remain similar across all NWP and the current criteria.

REFERENCES

- Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.*, **517**, 913–922, <https://doi.org/10.1016/j.jhydrol.2014.06.035>.
- , B. Bisselink, F. Dottori, G. Naumann, A. de Roo, P. Salamon, K. Wyser, and L. Feyen, 2017: Global projections of river flood risk in a warmer world. *Earth's Future*, **5**, 171–182, <https://doi.org/10.1002/2016EF000485>.
- Bartholmes, J., and E. Todini, 2005: Coupling meteorological and hydrological models for flood forecasting. *Hydrol. Earth Syst. Sci.*, **9**, 333–346, <https://doi.org/10.5194/hess-9-333-2005>.
- , J. Thielen, M. H. Ramos, and S. Gentilini, 2009: The European flood alert system EFAS—Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.*, **13**, 141–153, <https://doi.org/10.5194/hess-13-141-2009>.
- Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, <https://doi.org/10.1038/nature14956>.
- Bouttier, F., and H. Marchal, 2024: Probabilistic short-range forecasts of high-precipitation events: Optimal decision thresholds and predictability limits. *Nat. Hazards Earth Syst. Sci.*, **24**, 2793–2816, <https://doi.org/10.5194/nhess-24-2793-2024>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, <https://doi.org/10.1002/asl.170>.
- Burek, P., J. van der Knijff, and A. de Roo, 2013: LISFLOOD, distributed water balance and flood simulation model. Joint Research Centre - European Commission Revised User Manual, 150 pp., <https://doi.org/10.2788/24719>.
- Caretta, M., and Coauthors, 2023: Water. *Climate Change 2022—Impacts, Adaptation and Vulnerability*, H.-O. Pörtner et al., Eds., Cambridge University Press, 551–712, <https://doi.org/10.1017/9781009325844.006>.
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, <https://doi.org/10.1016/j.jhydrol.2009.06.005>.
- , —, P. J. Smith, and F. Wetterhall, 2017: How do I know if I've improved my continental scale flood early warning system? *Environ. Res. Lett.*, **12**, 044006, <https://doi.org/10.1088/1748-9326/aa625a>.
- Dasgupta, A., and Coauthors, 2025: Connecting hydrological modeling and forecasting from global to local scales: Perspectives from an international joint virtual workshop. *J. Flood Risk Manage.*, **18**, e12880, <https://doi.org/10.1111/jfr3.12880>.

- Demeritt, D., S. Nobert, H. L. Cloke, and F. Pappenberger, 2013: The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Processes*, **27**, 147–157, <https://doi.org/10.1002/hyp.9419>.
- De Roo, A. P. J., C. G. Wesseling, and W. P. A. Van Deursen, 2000: Physically based river basin modelling within a GIS: The LISFLOOD model. *Hydrol. Processes*, **14**, 1981–1992, [https://doi.org/10.1002/1099-1085\(20000815/30\)14:11/12<1981::AID-HYP49>3.0.CO;2-F](https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F).
- , and Coauthors, 2003: Development of a European flood forecasting system. *Int. J. River Basin Manage.*, **1**, 49–59, <https://doi.org/10.1080/15715124.2003.9635192>.
- Diomedea, T., and Coauthors, 2008: Discharge prediction based on multi-model precipitation forecasts. *Meteor. Atmos. Phys.*, **101**, 245–265, <https://doi.org/10.1007/s00703-007-0285-0>.
- EEA, 2023: Economic losses from weather- and climate-related extremes in Europe. European Environmental Agency - European Union, <https://www.eea.europa.eu/en/analysis/indicators/economic-losses-from-climate-related?activeAccordion=ecd3b3cf-bbe9-4978-b5cf-0b136399d9f8>.
- Gouweleeuw, B. T., J. Thielen, G. Franchello, A. P. J. de Roo, and R. Buizza, 2005: Flood forecasting using medium-range probabilistic weather prediction. *Hydrol. Earth Syst. Sci.*, **9**, 365–380, <https://doi.org/10.5194/hess-9-365-2005>.
- Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- Hanssen, A. W., and W. J. A. Kuipers, 1965: *On the Relationship between the Frequency of Rain and Various Meteorological Parameters*. Mededelingen en Verhandelingen—Koninklijk Nederlands Meteorologisch Instituut, Vol. 81, Staatsdrukkerij- en Uitgeverijbedrijf, 65 pp.
- JRC, 2020: European Flood Awareness System v4.0. Joint Research Centre - European Commission, accessed 1 October 2024, <https://confluence.ecmwf.int/display/CEMS/EFAS+v4.0>.
- , 2023: European Flood Awareness System v5.0. Joint Research Centre - European Commission, accessed 30 October 2024, <https://confluence.ecmwf.int/display/CEMS/EFAS+v5.0>.
- , 2024: European Flood Awareness System v5.2. Joint Research Centre - European Commission, accessed 30 October 2024, <https://confluence.ecmwf.int/display/CEMS/Latest+operational+EFAS+release>.
- Kling, H., M. Fuchs, and M. Paulin, 2012: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.*, **424–425**, 264–277, <https://doi.org/10.1016/j.jhydrol.2012.01.011>.
- Knoben, W. J. M., J. E. Freer, and R. A. Woods, 2019: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.*, **23**, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>.
- Lavers, D. A., and Coauthors, 2020: A vision for hydrological prediction. *Atmosphere*, **11**, 237, <https://doi.org/10.3390/atmos11030237>.
- LeClerc, J., and S. Joslyn, 2015: The cry wolf effect and weather-related decision making. *Risk Anal.*, **35**, 385–395, <https://doi.org/10.1111/risa.12336>.
- Legg, T. P., and K. R. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906, [https://doi.org/10.1175/1520-0434\(2004\)019<0891:EWOSWF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2).
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petrolagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, <https://doi.org/10.1002/qj.49712252905>.
- Nevo, S., and Coauthors, 2022: Flood forecasting with machine learning models in an operational framework. *Hydrol. Earth Syst. Sci.*, **26**, 4013–4032, <https://doi.org/10.5194/hess-26-4013-2022>.
- Pappenberger, F., K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. de Roo, 2005: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrol. Earth Syst. Sci.*, **9**, 381–393, <https://doi.org/10.5194/hess-9-381-2005>.
- , J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. de Roo, 2008a: New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.*, **35**, L10404, <https://doi.org/10.1029/2008GL033837>.
- , K. Scipal, and R. Buizza, 2008b: Hydrological aspects of meteorological verification. *Atmos. Sci. Lett.*, **9**, 43–52, <https://doi.org/10.1002/asl.171>.
- , M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.*, **522**, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>.
- Paprotny, D., A. Sebastian, O. Morales-Nápoles, and S. N. Jonkman, 2018: Trends in flood losses in Europe over the past 150 years. *Nat. Commun.*, **9**, 1985, <https://doi.org/10.1038/s41467-018-04253-1>.
- Ramos, M.-H., J. Bartholmes, and J. Thielen-del Pozo, 2007: Development of decision support products based on ensemble forecasts in the European flood alert system. *Atmos. Sci. Lett.*, **8**, 113–119, <https://doi.org/10.1002/asl.161>.
- Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, <https://doi.org/10.1175/2008WAF2222159.1>.
- Roulin, E., and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *J. Hydrometeorol.*, **6**, 729–744, <https://doi.org/10.1175/JHM436.1>.
- Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo, 2009a: The European Flood Alert System—Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13**, 125–140, <https://doi.org/10.5194/hess-13-125-2009>.
- , K. Bogner, F. Pappenberger, M. Kalas, M. del Medico, and A. de Roo, 2009b: Monthly-, medium-, and short-range flood warning: Testing the limits of predictability. *Meteor. Appl.*, **16**, 77–90, <https://doi.org/10.1002/met.140>.
- Wetterhall, F., and Coauthors, 2013: HESS Opinions “Forecaster priorities for improving probabilistic flood forecasts”. *Hydrol. Earth Syst. Sci.*, **17**, 4389–4399, <https://doi.org/10.5194/hess-17-4389-2013>.
- WMO, 2021: WMO Atlas of mortality and economic losses from weather, climate and water extremes (1970–2019). World Meteorological Organization Tech. Rep., 90 pp., https://library.wmo.int/viewer/57564/download?file=1267_Atlas_of_Mortality_en.pdf&type=pdf&navigator=1.
- Yang, S.-C., T.-H. Yang, Y.-C. Chang, C.-H. Chen, M.-Y. Lin, J.-Y. Ho, and K. Lee, 2020: Development of a hydrological ensemble prediction system to assist with decision-making for floods during typhoons. *Sustainability*, **12**, 4258, <https://doi.org/10.3390/su12104258>.