Received: 19 February 2020 | Revised: 28 May 2020 | Accepted: 1 July 2020

DOI: 10.1111/jfr3.12658

ORIGINAL ARTICLE

CIWEM Chartered Institution of Water and Environmental Management | Journal of Flood Risk Management | WILEY

Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting

Ervin Zsoter[1,2] | **Christel Prudhomme**[1,3,4] | **Elisabeth Stephens**[2] | **Florian Pappenberger**[1] | **Hannah Cloke**[2,5,6,7]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK

[2]Department of Geography and Environmental Science, University of Reading, Reading, UK

[3]UK Centre for Ecology & Hydrology, Wallingford, UK

[4]Department of Geography, Loughborough University, Loughborough, UK

[5]Department of Meteorology, University of Reading, Reading, UK

[6]Department of Earth Sciences, Uppsala University, Uppsala, Sweden

[7]Centre of Natural Hazards and Disaster Science, CNDS, Uppsala, Sweden

Correspondence
Ervin Zsoter, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK.
Email: ervin.zsoter@ecmwf.int

Funding information
Natural Environment Research Council, Grant/Award Numbers: NE/K00896X/1, NE/P000525/1

Abstract

Global flood forecasting systems rely on predefining flood thresholds to highlight potential upcoming flood events. Existing methods for flood threshold definition are often based on reanalysis datasets using a single threshold across all forecast lead times, such as in the Global Flood Awareness System. This leads to inconsistencies between how the extreme flood events are represented in the flood thresholds and the ensemble forecasts. This paper explores the potential benefits of using river flow ensemble reforecasts to generate flood thresholds that can deliver improved reliability and skill, increasing the confidence in the forecasts for humanitarian and civil protection partners. The choice of dataset and methods used to sample annual maxima in the threshold computation, both for reanalysis and reforecast, is analysed in terms of threshold magnitude, forecast reliability, and skill for different flood severity levels and lead times. The variability of threshold magnitudes, when estimated from the different annual maxima samples, can be extremely large, as can the subsequent impact on forecast skill. Reanalysis-based thresholds should only be used for the first few days, after which ensemble-reforecast-based thresholds, that vary with forecast lead time and can account for the forecast bias trends, provide more reliable and skilful flood forecasts.

KEYWORDS

ensemble reforecasts, flood forecasting, flood thresholds, forecast lead times, global predictions, reanalysis, river discharge

1 | INTRODUCTION

Flood forecasting systems use meteorological data and hydrological modelling to deliver forecasts of river discharge and other hydrological variables such as inundation or soil moisture. They provide early flood warnings on time scales up to several weeks ahead, essential for managing flood risk at local, regional, and recently also on the global scale (Emerton et al., 2016).

The state-of-the-art systems in use today provide an ensemble of equally likely solutions that can be used to define occurrence probabilities for certain flood events (Cloke & Pappenberger, 2009; Wu et al., 2020). These flood events are defined by comparing the forecast time

series with flood thresholds, usually based on a return period magnitude or a quantile.

In the Global Flood Awareness System of the Copernicus Emergency Management Service (GloFAS; Alfieri et al., 2013, Hirpa et al., 2018), the severity of the predicted flood is defined according to a set of three thresholds, as shown in Figure 1 for the example of tropical cyclone Idai in Mozambique in March 2019. These thresholds are computed from a 40-year long river discharge reanalysis (Harrigan et al., 2020). The hydrograph in Figure 1 shows the predicted river discharge for the next 30 days, highlighting a severe flood event around 18–21 March with 10–15% chance of exceeding the 5% annual exceedance probability (AEP) threshold.

The flood thresholds, defined according to flood magnitude of selected return periods (or flood quantiles), and used in many of the existing flood prediction systems (GloFAS Alfieri et al., 2013; EFAS Thielen, Bartholmes, Ramos, & de Roo, 2009; WW-HYPE Arheimer et al., 2020), are determined by flood frequency analysis, usually by fitting an extreme value distribution on a set of annual maxima, sampled from a time series as long as possible. These quantities describe the likelihood of different flood magnitudes occurring locally based on a "climatological" data set over a long period of time (preferably 30 years or more; World Meteorological Organisation [WMO], 2017). Traditionally, flood thresholds are produced from observations or deterministic model reanalysis (Alfieri et al., 2015). River discharge observations can provide a solution only at certain locations, whereas hydrological model simulations, forced with meteorological observations, can cover a whole geographical domain, delivering flood thresholds at every model river point or catchment.

Because the flood thresholds determine the severity of the forecasted flood signal, these flood thresholds should ideally represent extreme events the same way as they occur in the forecasts. If this is not the case and the different biases make an event of the same magnitude occur with a different frequency in the climatological data set that was used to compute the thresholds and in the forecasts (e.g., the 5% AEP flood magnitude happens more often in the forecasts than the expected 5% probability in a given year), then the flood forecast probabilities could become unreliable (e.g., leading to flood signals that often overestimate the flood severity). In the case of the

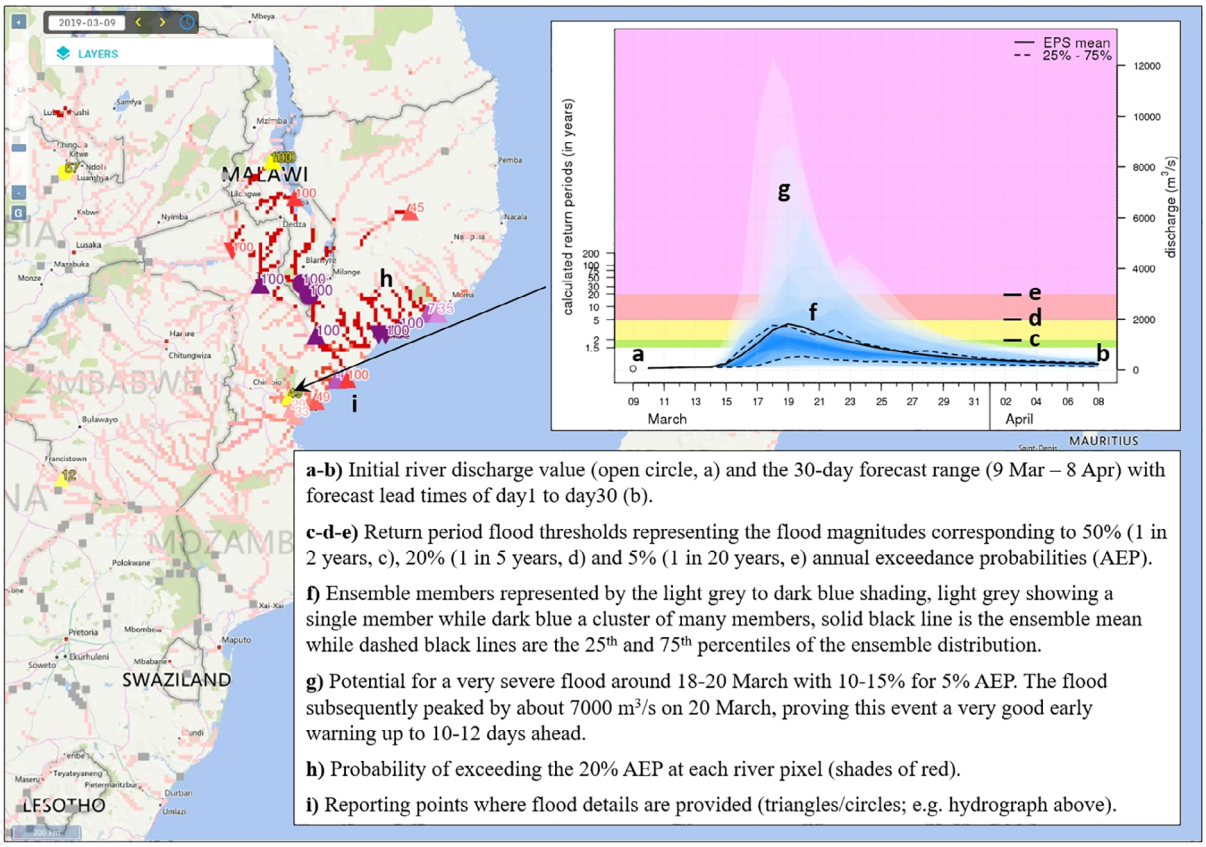


FIGURE 1 GloFAS forecast on March 9, 2019 for Mozambique showing flood predictions related to tropical cyclone Idai. As an example, the inset diagram shows the hydrograph for a river point near the coast in Mozambique for the 30-day period of 9 March to 8 April. GloFAS forecasts are openly accessible on www.globalfloods.eu

example in Figure 1, this could mean that the predicted severe flood event should in fact appear significantly less extreme as the high severity would only be a consequence of the unrealistically low thresholds.

The extreme event representation of flood thresholds can be heavily influenced by the data set and the method used to derive the thresholds. The value of the flood quantiles can be impacted by the choice of the extreme value statistical distribution (Papalexiou & Koutsoyiannis, 2013), the data set that is used for the annual maxima extraction (observation, reanalysis or forecasts; see, for example, Hirpa et al., 2016) and also its length (Kjeldsen, Lamb, & Blazkova, 2014). These can all lead to potential differences in the flood threshold magnitudes, subsequently resulting in differences in the forecast probabilities to exceed the thresholds, and ultimately causing an impact on the quality of the flood warnings.

By definition, conventional observation- or reanalysis-based data sets provide a single time series to compute flood thresholds, meaning that only one set of thresholds, with different severities, is going to be applied to all lead times in the forecast range. This might cause further inconsistencies if the forecast biases have trends across lead times. For example, forecasts might show increasing river discharge overprediction with lead time, which would result in a growing number of forecasts exceeding the 5% AEP flood threshold (which stays unchanged as it is computed from the reanalysis time series), with an increasingly higher frequency than the expected 5% of the years occurrence on average (Alfieri et al., 2019). As trends and biases in a forecast are model specific, using different meteorological forcing models within the same forecasting system (such as in the European Flood Awareness System, EFAS, Thielen, Bartholmes, Ramos, & de Roo, 2009) might cause even more complex inconsistencies between the observation- or reanalysis-based flood thresholds and the forecasts.

Bias correction methods can help to achieve consistency between forecasts and thresholds (e.g., Verkade, Brown, Reggiani, & Weerts, 2013; Yuan & Wood, 2012). They have the potential to make the extreme event representation of the forecasts and the climatology, that is used to define the thresholds, similar. However, bias correction, even in its simplest form with only hydrological output postprocessing (without correction of the meteorological forcing data), would introduce further complexity into the river discharge production chain with its associated uncertainties.

Alternative approaches have been investigated (e.g., Alfieri et al., 2019). Generally, flood thresholds are not produced from forecasts. Part of the reason could be the limited sample of available historical forecasts and also the convenience for the users to work with only a single threshold set that does not show evolution with lead time. The consequence is that, as said earlier, the same threshold is applied to all forecast lead times. However, Alfieri et al. (2019) showed that range-dependent, reforecast-based thresholds were substantially different from unique reanalysis-based thresholds in two thirds of the global rivers. Moreover, despite the recent advancement of ensemble-based forecast systems, ensemble forecasts are generally not considered in the flood threshold generation. However, this can be a problem as ensembles can have different biases to single deterministic forecasts (Leutbecher et al., 2017), which can further contribute to the extreme event representation inconsistencies between reanalysis-based thresholds and ensemble forecasts.

The use of ensemble reforecasts in generating the climatological sample can provide a range-dependent threshold set (e.g., as in Emerton et al., 2018 and Tsonevsky, Doswell, & Brooks, 2018), which has the potential to overcome the issues associated with extreme event identification. In addition, multi-value ensembles can also contribute to increased effective sample size, from which to define flood thresholds, and therefore help to improve the representation of extreme events (Zsoter, Pappenberger, & Richardson, 2014). This could be important for very extreme events which might not occur in the typical 30–50-year-long sample of traditional observation or reanalysis time series (e.g., the median length of the daily data in the Global Runoff Data Centre is 39 years, as of January 22, 2020 at www.bafg.de/GRDC).

In this study, the potential benefits of using river discharge ensemble reforecasts to define flood thresholds are analysed globally. Two main research questions were explored in our study, targeting specifically the sampling strategy to extract the annual maxima sample on which the flood frequency analysis is conducted:

- How adequate it is to use a reanalysis dataset to define flood thresholds and apply them for all forecast lead times?
- How best to use reforecast ensemble information in the flood threshold generation to improve flood forecast performance?

The work is carried out in the context of GloFAS, for a 30-day forecast range, with a selection of over 5,000 catchments. The impacts of the choice of data source (reanalysis or reforecasts) and of the annual maximum sampling strategies (from the reforecasts) are analysed by comparing the flood threshold magnitudes and the resulting forecast reliability and skill benefits for four different flood severity levels.

2 | SYSTEM DESCRIPTION, DATASETS, AND METHODS

This section describes the data sets, methods, and experimental set-up used to generate flood thresholds and compare their value and impact on the flood forecast skill.

2.1 | GloFAS

GloFAS is part of the Copernicus Emergency Management Service (CEMS) and has been developed by the Joint Research Centre of the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF) with help from research institutions such as the University of Reading (UoR; e.g., Stephens, Day, Pappenberger, & Cloke, 2015, Emerton, Cloke, & Stephens, 2017 and Towner et al., 2019). It is a probabilistic hydrological prediction system, which has a 30-day (Alfieri et al., 2013) and a seasonal component (Emerton et al., 2018). This study is based on the 30-day component, which predicts daily flood occurrences on the global scale. In GloFAS, ensemble runoff outputs from the HTESSEL land surface model (the Hydrology-Tiled ECMWF Scheme for Surface Exchange over Land; Balsamo et al., 2009; Balsamo, Pappenberger, Dutra, Viterbo, & van den Hurk, 2011) are coupled to the Lisflood hydrological model (van der Knijff, Younis, & de Roo, 2010) to produce an ensemble of daily river discharge across a global river network at 0.1° resolution (Alfieri et al., 2013; Hirpa et al., 2018). To detect the likelihood of high flow situations, to forecast flood events, the real time river discharge forecasts are compared with a set of flood thresholds derived from a 40-year long climatological simulation, a daily river discharge reanalysis time series.

2.2 | River discharge reanalysis

The GloFAS-ERA5 river discharge reanalysis (Harrigan et al., 2020) is produced with ERA5 forcing, ECMWF's fifth generation global climate reanalysis (Hersbach et al., 2018; Hersbach et al., 2020), which is part of the EU-funded Copernicus Climate Change Service (C3S). ERA5 covers the period 1979 to present and is updated with two to 3 months delay. ERA5 is open access (https://climate.copernicus.eu/) and includes one high-resolution component and a lower resolution ensemble component with 10 members. The GloFAS-ERA5 uses the high-resolution ERA5 component at ∼31 km horizontal resolution with the configuration of the GloFAS operational forecasting systems. GloFAS-ERA5 is a key component of GloFAS verification, serving as a proxy for river discharge observations and it is also openly available from the Copernicus Climate Change Service Climate Data Store (Harrigan et al., 2020).

2.3 | Ensemble river discharge reforecasts

The ensemble river discharge reforecasts are GloFAS reforecasts produced for the 20-year period of 1997–2016. These are 30-day river discharge forecasts generated for past dates by the same GloFAS system that is used for the real time forecasts. They are initialised from GloFAS-ERA5 and forced by runoff from the twice weekly (Monday and Thursdays in 2017), 11-member, 20-year ECMWF meteorological ensemble reforecasts (Vitart, 2014). This data set includes a batch of 20 reforecasts (one for each year in 1997–2016) for each Monday and Thursday in 2017. Altogether 2080, 11-member, 30-day reforecasts were produced for the 20-year period (104 in each year).

2.4 | Flood thresholds

In the 30-day GloFAS, flood quantiles of three severity levels (2-, 5-, and 20-year return periods) are used as flood thresholds. Flood quantiles are commonly used in risk analysis, typically estimated using time series data of generally twice the length of the return period of interest. Because of the relatively short length of daily discharge data available, the 10-year return period severity was also considered in this study.

A return period T is an estimate of the likelihood of an event to occur (Gumbel, 1941), expressed as average number of years for an event of same or higher magnitude to occur. It can also be expressed as an Annual Exceedance Probability AEP (given by $\mathrm{AEP} = 100/T$). To facilitate the interpretation, AEP is used in the rest of this study.

The flood thresholds were computed as currently done operationally in GloFAS: the Gumbel Extreme Value Distribution (EVD) is fitted to the annual maximum river discharge sample using the method of L-Moments (Hosking, 1990). This method is appropriate for relatively small sample sizes, such as used in GloFAS (Alfieri et al., 2019) and in this study (20 years of reforecast data from 1997 to 2016).

2.5 | River catchments

The study is based on the GloFAS network (Figure 2), a set of 6,122 catchments of which about

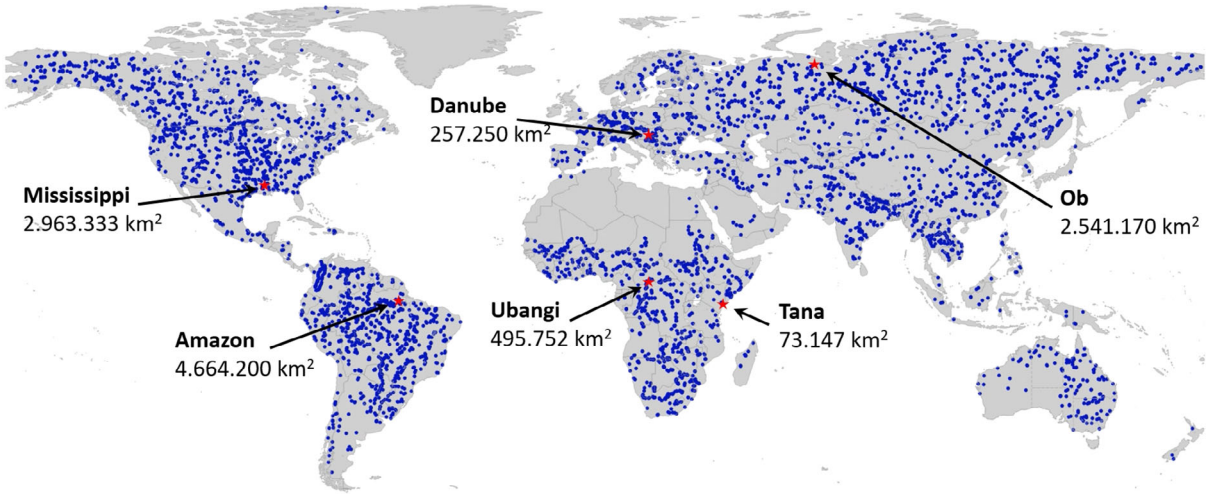
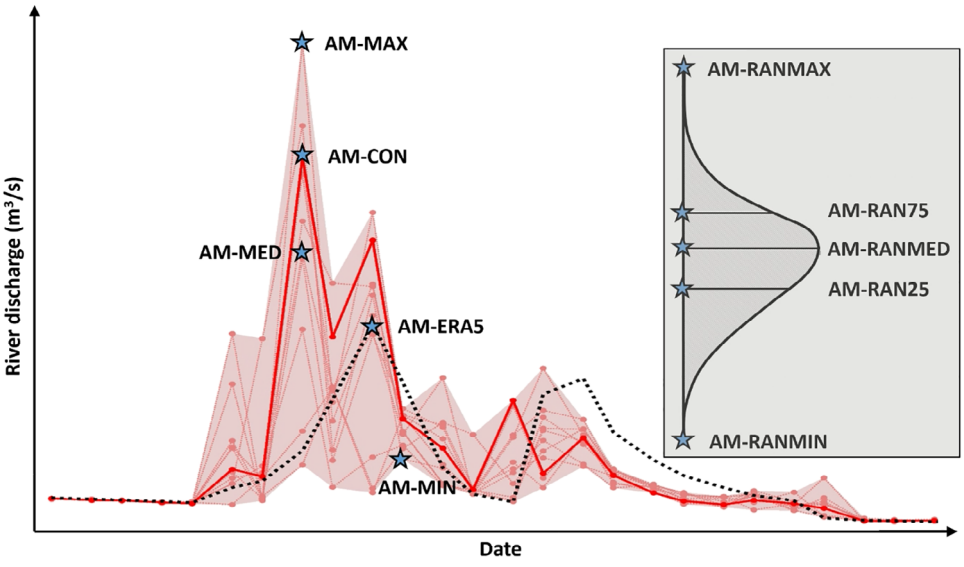


FIGURE 2 The 5,665 GloFAS stations used in this study. The six contrasting catchments of Figure 5 are indicated by red stars, along with the river names and GloFAS upstream areas



Version	Time series set	Version	Time series set
T-ERA5	Benchmark set	T-RANMIN	Minimum extended reforecast set
T-CON	Control member reforecast set	T-RAN25	25th percentile extended reforecast set
T-MIN	Minimum member reforecast set	T-RANMED	Median extended reforecast set
T-MED	Median member reforecast set	T-RAN75	75th percentile extended reforecast set
T-MAX	Maximum member reforecast set	T-RANMAX	Maximum extended reforecast set

FIGURE 3 Schematic of the annual maximum sampling for flood threshold estimates from daily river discharge time series. Dotted black line: GloFAS-ERA5, solid red line: GloFAS reforecast control member, light red lines: GloFAS reforecast perturbed ensemble members. Small red dots show individual daily river discharge values in the ensemble reforecasts. The *x*-axis shows the date of the forecasts, while the *y*-axis the river discharge values

one-third are always highlighted on GloFAS website as reporting points (www.globalfloods.eu). This network provides a global coverage and includes all points where daily historical river discharge observations are made available to the GloFAS team. Catchments that have 50% AEP magnitude below 20 m^3/s in GloFAS-ERA5, that is, too dry or too small, were excluded from the study, resulting in 5,665 catchments in total for the analysis.

2.6 | Analysis methods

The impact of flood threshold estimation for all four severity levels was analysed for each catchment by direct comparison of the quantile magnitudes. Additionally, the flood forecast performance was evaluated for day 1 to day 30 lead times by:

- Comparing the number of events forecasted (i.e., when the discharge exceeds the flood threshold) with the

number of events identified in the benchmark set (i.e., GloFAS-ERA5 river discharge reanalysis which is the nearest equivalent to the "observations"), expressed as percentage occurrence frequency (or event/forecast probability). This step analyses the simplified forecast reliability with only one probability category;
- Calculating the Brier score (Murphy, 1973) and the reliability diagram (Hsu & Murphy, 1986). This step assesses both the skill and reliability in the resulting probability forecasts of exceeding the thresholds.

2.7 | Experimental set-up

For consistency and comparability, the annual maxima sampling for the flood threshold computation was done from daily time series containing only the calendar days corresponding to the dates of the day 1 to day 30 reforecast values (for all Monday and Thursday reforecast runs of 1997–2016). For each lead time, three sets of time series were used:

- *Benchmark set*: GloFAS-ERA5 river discharge reanalysis (independent from the lead time). This is as close as possible to the flood thresholds used operationally in GloFAS and can be considered as proxy observation-based thresholds;
- *Reforecast set*: the time series of the control member, plus three time series corresponding to the minimum, median and maximum values from each run of the 11-member GloFAS reforecasts;
- *Extended reforecast set*: 1,000 time series, each generated with randomly selecting one of the 11 ensemble members from each GloFAS reforecast.

After applying the flood threshold generation method, described earlier, this resulted in 1 + 4 + 1,000 threshold values summarised graphically in Figure 3 for the annual maxima selection differences. From the 1,000 random-member-based thresholds only the minimum, 25th percentile, median, 75th percentile, and maximum values were analysed further. The exercise was conducted on all study catchments, flood severity levels, and forecast lead times. The major methodological steps of this study are provided in Table 1.

TABLE 1 Major methodological steps of this study

Steps	Description
Setup	Ensemble reforecasts for day 1 to day 30 lead times, over 20 years (1997–2016), with 104 forecasts in each year, flood thresholds computed by fitting an extreme value distribution on the 20 annual maxima, for 5,665 global catchments and 4 return periods (50, 20, 10, and 5% AEP)
Benchmark (reanalysis) thresholds	Produce reanalysis-based reference thresholds (T-ERA5) for all lead times, always with the days of the reforecasts, to guarantee homogeneous samples
Reforecast thresholds	Produce ensemble-reforecast-based alternative thresholds (T-CON, T-MIN, T-MED, T-MAX)
Extended reforecast thresholds	Produce random-ensemble-member-based thresholds 1,000 times (T-RAN)
Extended reforecast threshold distribution	Define the key statistics of the extended reforecast threshold distribution (T-RANMIN, T-RAN25, T-RANMED, T-RAN75, and T-RANMAX)
Probabilities	Compute the exceedance probabilities over the 20-year period with all threshold versions (10 in total)
Scores	Compute the Brier scores and produce the reliability diagrams with all threshold versions (10 in total), for all catchments including a global average

Note: For all lead times, catchments and return periods if not otherwise stated.

3 | RESULTS

The impact of the data set and sampling strategy choice in the flood threshold generation was analysed globally on selected river catchments, with the flood threshold magnitude and forecast skill compared geographically.

3.1 | How similar are the flood thresholds?

In this section, we analyse the impact of the annual maxima sampling strategy on the flood threshold magnitude for day 1 and day 30 lead times, focusing on the 10% AEP severity level.

Flood threshold magnitudes, derived from reforecasts, depend on lead-time with values less than 5% different from those derived from the reanalysis (T-ERA5) for day 1 (Figure 4a,c,e), but exceeding 50% difference over large parts of the world by day 30 (Figure 4b,d,f), regardless of the ensemble reforecast sampling strategy. However, there is a large spatial

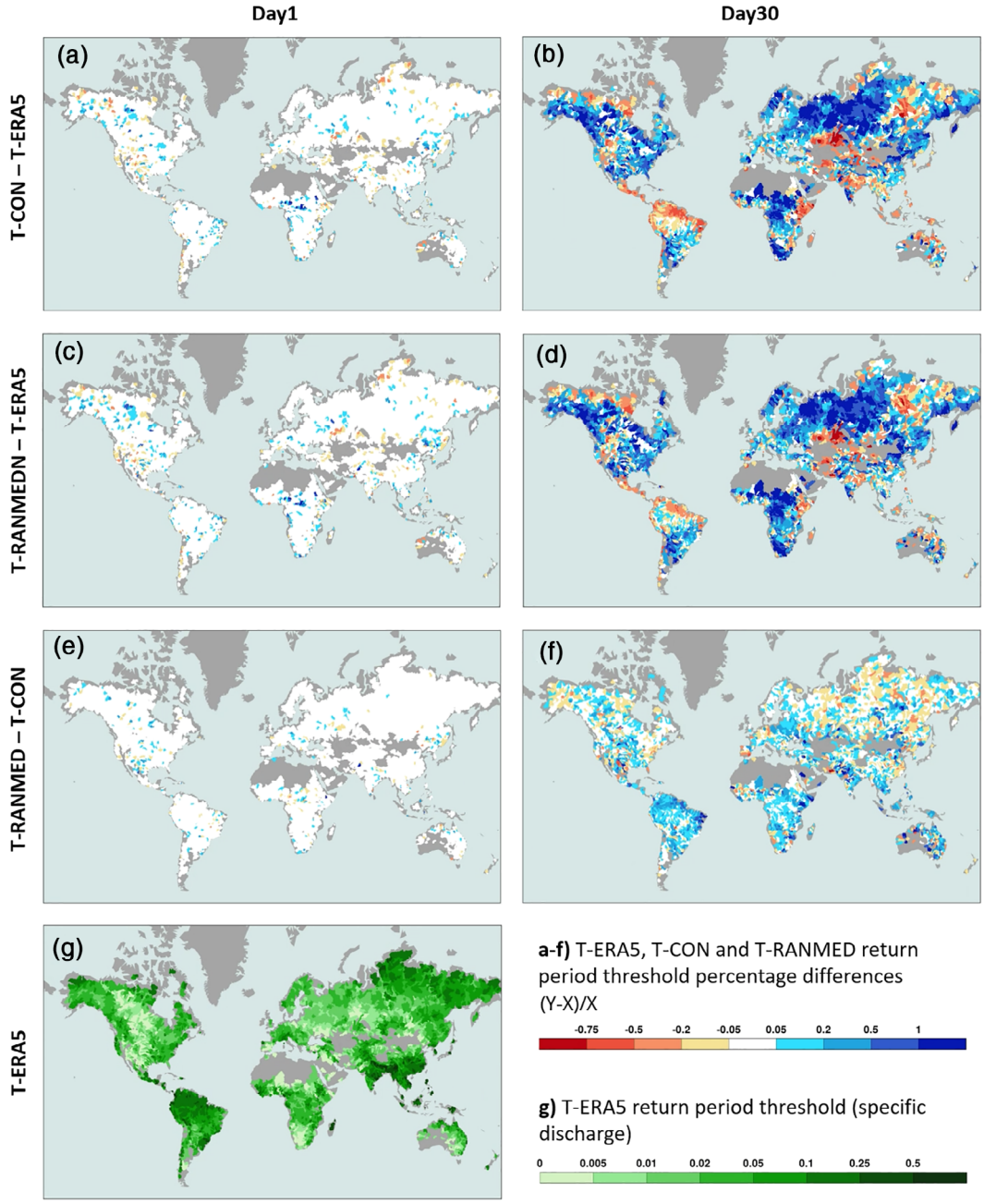


FIGURE 4 Percentage difference of 10% AEP flood thresholds between (a, b) T-CON and T-ERA5, (c, d) T-RANMED and T-ERA5 and (e, f) T-RANMED and T-CON based on the 1997–2016 period. The left column (a, c, e) is for day 1 while the right one (b, d, f) is for day 30 lead time. Percentage differences of orange (blue) colour palette mean lower (higher) flood thresholds respectively in T-CON (vs. T-ERA5) and in T-RANMED (vs. T-ERA5 and T-CON). Panel (g) shows the reference T-ERA5 threshold magnitudes as specific river discharge (river discharge divided by the upstream area in km^2 in order to scale better between different catchment sizes)

variability, with most of the world showing reforecast flood thresholds larger than T-ERA5, except in north Canada, Central and northern South America, Central Asia, the Horn of Arica, and some of east Russia. This confirms earlier finding from Alfieri et al. (2019) that for extended-range lead times, the flood frequency distribution of hydrological forecasts is not well represented by reanalysis simulations.

Figure 4 also shows that using the control member of the reforecasts to derive the flood threshold (T-CON) leads to different results than sampling the full reforecast ensemble (T-RAN family). Specifically, T-CON is systematically higher than T-RANMED in most regions (Figure 4e,f), with differences growing with lead time, generally below 20% but reaching 100% in some places by day 30. This suggests that the control

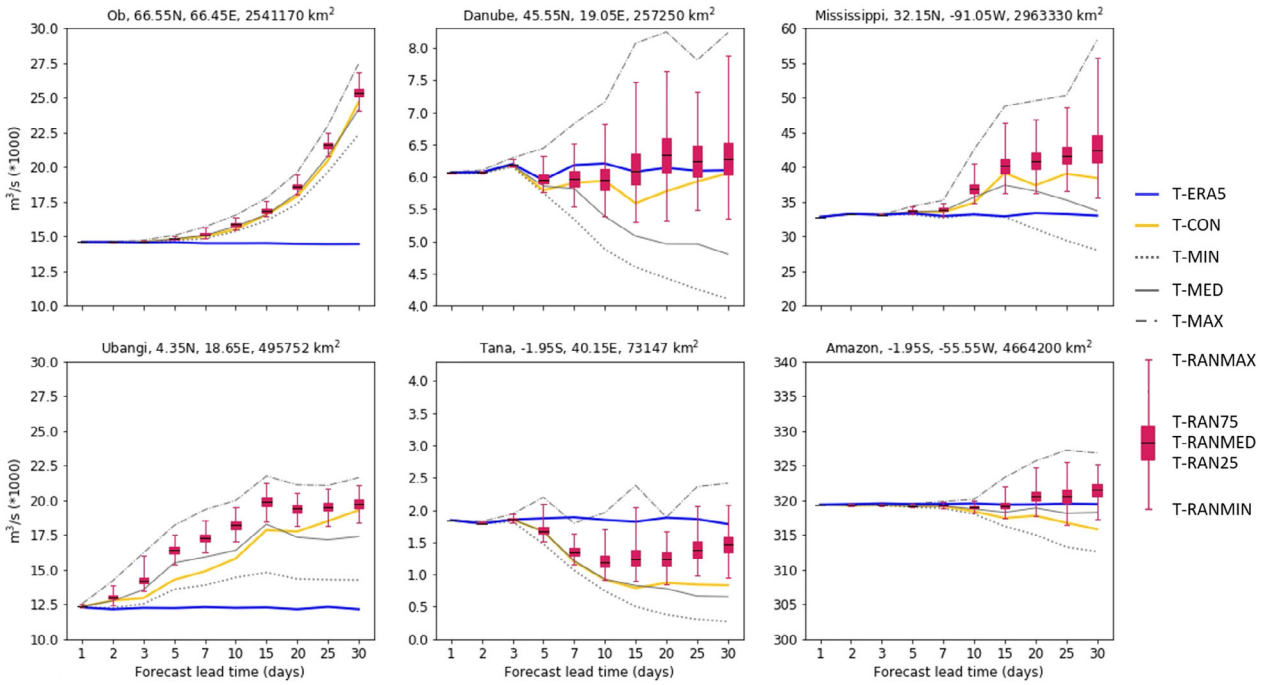


FIGURE 5 Flood thresholds of 10% AEP severity level based on the 1997–2016 period as function of forecast lead time for six contrasting catchments: benchmark T-ERA5 (blue), reforecast control T-CON (orange), minimum T-MIN (dotted grey), median T-MED (solid grey) and maximum T-MAX (dash-dotted grey) and also the extended reforecast T-RANMIN, T-RAN25, T-RANMED, T-RAN75, and T-RANMAX (red box whiskers)

forecasts do not fully represent the flood frequency distribution of the ensemble reforecasts and their use could potentially lower the forecast skill.

Results for other severity levels (5, 20, and 50% AEPs) show very similar behaviour. Although there are some variabilities across the severity levels with differences in flood threshold magnitude increasing with the severity level, the percentage differences appear to be in the same order of magnitude (Figures S1–S3).

Figure 5 shows the flood thresholds for the 10% AEP severity level as a function of lead time for six contrasting catchments (consult Figure 2 for the catchment locations). The influence of the ensemble reforecast sampling strategy on the flood threshold magnitudes gets larger with the increasing forecast lead time. For some catchments, such as the Ob and Amazon rivers, the impact is small (interquartile range of below 1% of T-ERA5 by day 30 as shown by the red boxes), but for some other catchments the difference could be as large as 10–20% of the T-ERA5 value at day 30 lead time (Tana and Mississippi rivers). Moreover, the flood thresholds, generated using the control member, are dominantly below the envelope of the ensemble reforecast (e.g., Ubangi and Tana rivers), confirming the general positive pattern already seen in Figure 4f.

Analysis on other flood threshold severity levels indicates that differences between reforecast- and reanalysis-based thresholds and sampling strategies are generally increasing with both severity level and lead time (Figures S4–S6).

3.2 | How reliable are the forecast probabilities?

In this section, we investigate the match between the flood forecast probabilities and the flood occurrence frequencies, using the benchmark, the reforecast control, and the extended reforecast median flood thresholds, defined for the 10% AEP severity level (Figure 6).

At day 1 (Figure 6a–c), flood forecasts are very reliable regardless of the flood threshold generation used (points close to the diagonal line), but this is lost by day 30 (Figure 6d–f). The largest loss of reliability is found when using the benchmark flood threshold (T-ERA5), with many catchments showing too high flood forecast probability (points way above the diagonal line), suggesting that the T-ERA5 thresholds are too low. The performance using reforecasts-based thresholds shows a clear improvement over using T-ERA5, especially reducing the number of catchments with large flood forecast probability overestimation. Results based on T-RANMED are slightly better than those using T-CON with a larger cluster around the diagonal line (91 vs. 86% of the

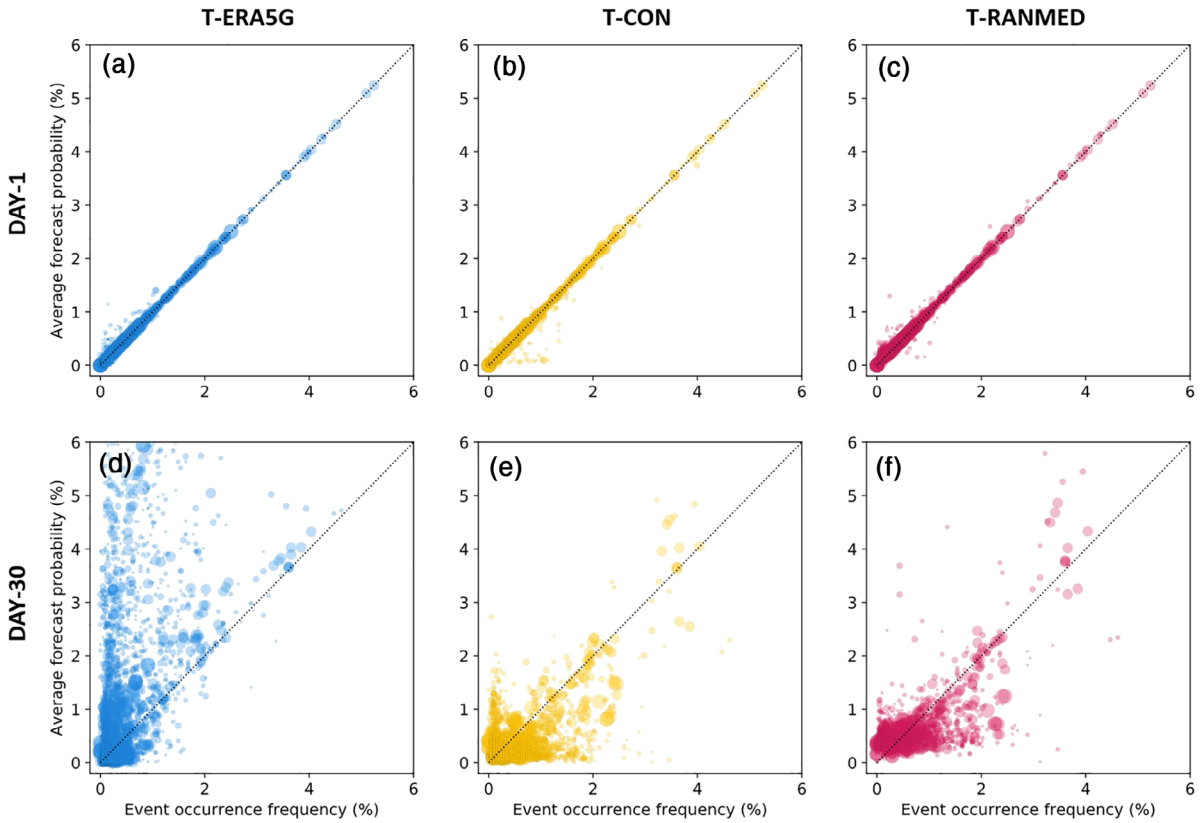


FIGURE 6 Scatter plot of day 1 (top) and day 30 (bottom) flood forecast probability (*y*-axis) against flood occurrence frequency (*x*-axis) using flood thresholds of T-ERA5 (a and c, blue), T-CON (b and e, orange), and T-RANMED (c and f, red) based on the 1997–2016 period. Dot size is proportional to catchment size

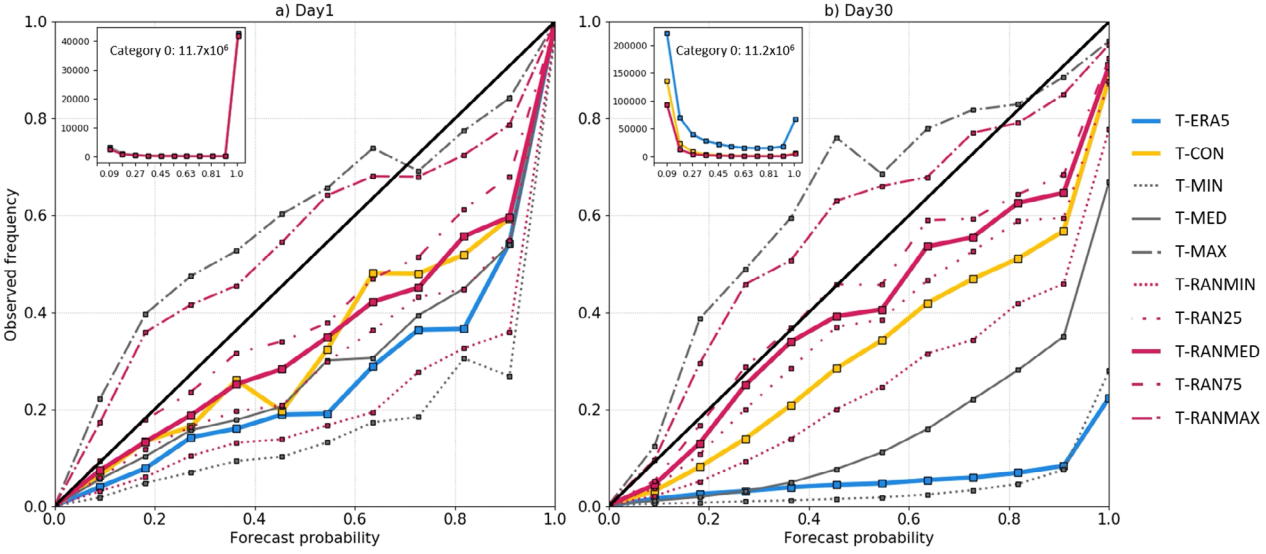


FIGURE 7 Reliability diagram for flood event forecast probabilities above 10% AEP based on the 1997–2016 period for (a) day 1 and (b) day 30 using flood thresholds based on the benchmark (T-ERA5), reforecast (T-CON, T-MIN, T-MED, and T-MAX), and extended reforecast (T-RANMIN, T-RAN25, T-RANMED, T-RAN75, and T-RANMAX) sets. The inset shows the distribution of number of cases in all 11 probability categories. The first category (0 ensembles member forecasting the event) is only indicated as a number

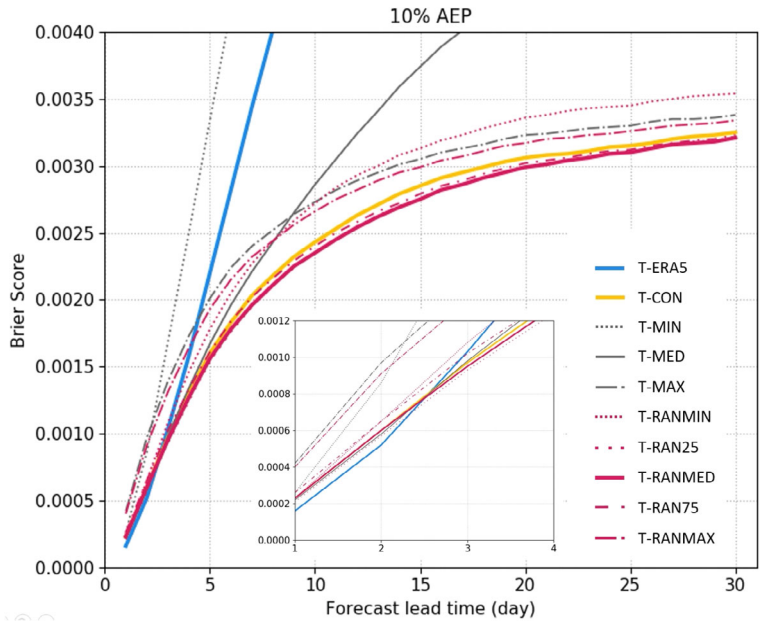


FIGURE 8 Brier score for flood event forecasts above 10% AEP over the 1997–2016 period for day 1 to day 30 using the benchmark (T-ERA5), reforecast (T-CON, T-MIN, T-MED, and T-MAX), and extended reforecast (T-RANMIN, T-RAN25, T-RANMED, T-RAN75, and T-RANMAX) flood thresholds. The inset shows the scores of the first 4 days only for better readability

catchments with less than 0.5% absolute difference between forecast probability and occurrence frequency), showing a stronger, more linear relationship.

3.3 | What is the impact on forecast reliability and skill?

Forecast reliability and skill were further examined using the reliability diagram and the Brier score for flood forecasts produced with the 10% AEP severity level.

As suggested by Figure 7, the reliability of the flood forecasts, based on the T-ERA5 thresholds, is low, especially for day 30 lead time, with an event frequency of less than 10% for almost all flood forecast probability categories (except the largest when it is just above 20%, see blue line close to *x*-axis). Using reforecast-based flood thresholds can greatly improve the flood forecast reliability, the only exceptions being T-MIN and also T-RANMIN for day 1 lead time. T-MAX and T-RANMAX tend to systematically underestimate flood event frequency up to 70–80% forecast probability, whilst overestimation of flood events is systematic for all other thresholds. The response is similar across all considered four severity levels except for flood thresholds of 50% AEP, where the reforecast thresholds become too high, making the flood forecast probabilities too low (points are above the diagonal; Figures S7–S9). Generally, thresholds based on the full ensemble can provide better reliability than using T-CON, this is especially clear by day 30, when all T-RANMED, T-RAN25, and T-RAN75 are closer to the diagonal line.

Figure 8 shows the general skill of the flood forecasts for the 10% AEP severity level from day 1 to day 30 lead times, based on the Brier score. The benchmark flood thresholds (T-ERA5) can provide the lowest error only up to day 2 lead time (blue line below the other lines; for other flood severity levels this maximum lead time ranges from day 1 (5% AEP) to day 4 (50% AEP), Figures S10–S12). From day 3, the flood forecasts with reforecast-based thresholds become gradually more skilful than with T-ERA5 (the only exception is T-MIN), consistently with the conclusions of Figure 7b. In fact, both T-ERA5 and T-MIN show poor skill with multiple times higher Brier score values by day 30 (this higher section of the Brier score range is not shown in Figure 8 for better readability). The best performance is achieved by the median (T-RANMED) and the interquartile range boundaries (T-RAN25 and T-RAN25) of the extended reforecast threshold set and the reforecast control thresholds (T-CON), with skill slowly degrading with lead time. The skill improvement, using T-RANMED over T-CON, is statistically significant at the 99% level from day 5–6 lead time (tested by bootstrapping the dates in the verification sample, Figure S13). The pattern is similar for flood events of higher severity (5% AEP), whilst for less severe floods (20 and 50% AEP), the highest skill is achieved using T-RAN25 or T-RANMIN, but T-RANMED is still achieving high skill (Figures S10–S12).

4 | DISCUSSION

The global analysis conducted here showed that using flood thresholds based on reforecasts improved substantially the forecast performance after the first 1–4 days of the forecast range (depending on the flood severity levels)

compared with using thresholds based on reanalysis, as done operationally in most forecast systems. One of the key advantages is the lead time specific definition of thresholds, which accounts for the changing representation of extreme event frequencies in the forecasts. Overall, forecast errors are reduced by up to 2–4 times compared with reanalysis-based thresholds, depending on the flood severity and lead time. Results also showed that using the single unperturbed control member to define the thresholds is not sufficient and exploring the full ensembles of the reforecasts in the threshold derivation further increases the forecast reliability and skill.

4.1 | Ensemble member independence

Using ensemble members allows a better representation of the extreme events in the forecast climatology by increasing the sample size. The ensemble members are correlated to some extent by sharing the same initial condition, especially at the beginning of the 30-day forecast horizon. Correlation between ensemble members reduces with increased lead time when each ensemble member drifts towards becoming an independent and identical random sample from the mode climate. In addition, for each of the reforecast-based threshold methods, only one member was chosen from all the twice weekly reforecasts in the 20-year period in order to increase independence. This guaranteed that the correlation between the individual reforecast values in the climatological sample remained very small. This made them an effectively independent realisation of the true underlying model climate distribution, ultimately providing an appropriate basis for the extreme value distribution fitting in the flood threshold computation.

4.2 | Best performing thresholds

The median of the extended reforecast threshold set, produced by using one random ensemble member from the reforecasts, provides the best overall performance, however, for lower severity levels some other reforecast-based thresholds can be slightly better. This can be related to the nonlinear response between reforecast ensemble time series and flood quantile estimation. In particular, with increasing lead time, outliers associated with very high forecasted river discharge become more likely within the 11 ensemble members. The annual maximum selection then will over-represent the high outliers through the random member selection process, as even if only one very high forecast value is selected in 1 or 2 years, it is likely to shift the estimate of the 5–10% AEP flood quantile to a high value. This potential increase of flood threshold value with lead time does not affect the forecast probabilities of flood event occurrence to the same extent, as the probabilities are calculated considering the full ensemble and are influenced much less by these relatively rare outliers in some of the reforecast members. This different effect of outliers on flood thresholds and flood forecast probabilities will translate into inconsistent reliability and skill impact associated with the various ways to sample the reforecasts to produce flood threshold, and could result, in some cases, in favouring a different sampling strategy than picking up the median.

4.3 | Biases in the forecasts

Using range-dependent flood thresholds, based on ensemble reforecasts, can account for the evolving biases in the forecasts across the forecast range. This study demonstrated that biases can grow large, affecting the extreme event representation and the use of flood thresholds in medium to extended range hydrological forecast systems like GloFAS 30-day. These biases can originate from the meteorological forcing and impact the hydrological simulations, mainly through precipitation and marginally also temperature, humidity, wind, and radiation, as shown by Zsoter et al. (2016) for the first 10 days of the forecasts. Another likely source for the biases is the land data assimilation (LDAS) impact documented by Zsoter et al. (2019). The LDAS can result in not conserving the water budget in coupled land surface models such as used in GloFAS, possibly contributing to biases seen in the GloFAS-ERA5 reanalysis across large parts of the world (Harrigan et al., 2020). In GloFAS, the reforecasts are initialised from GloFAS-ERA5, but with increasing lead time, the influence of LDAS on reforecast gradually decreases. This means that biases coming from the LDAS impact will remain present in any reanalysis-based flood threshold (in our case GloFAS-ERA5) but will slowly disappear with lead time in reforecasts-based flood thresholds. This inconsistency is likely to contribute to the large differences between the GloFAS-ERA5- and the ensemble-reforecast-based thresholds shown in this study.

4.4 | Forecast post-processing

Post-processing of the forecasts against the reference dataset used to derive the flood thresholds (i.e., in our case, GloFAS-ERA5) is an alternative to ensemble reforecast-based thresholds. By removing biases in the forecasts (e.g., linear regression or quantile mapping; see Wentao et al., 2017 for a review of methods), the extreme

event representation of the forecasts would be expected to become similar to that of the reference dataset or climatology. The use of post-processing techniques to create a consistent system between forecasts and flood thresholds was beyond the scope of this paper but could be pursued in the future.

4.5 | Modelling system independence

Whilst the research was conducted on the GloFAS flood forecasting system (based on the HTESSEL land surface model), the main findings of this work are expected to be independent of the modelling system, the extreme value fitting method or the sampling period length used. Although a different fitting method or sampling length could inevitably change the flood thresholds locally, in the global context, they are expected to have a neutral impact on the relation of the threshold magnitudes amongst the different annual maxima sampling methods. In addition, the forecast biases are bound to be modelling system related, which will inevitably change the flood threshold behaviour across the forecast lead times. However, the benefit of using ensemble-based, lead-time-specific thresholds is expected to be general and not dependent on the actual underlying bias behaviour. This is supported by the consistent results found using the Lisflood hydrological model in Alfieri et al. (2019), where the reforecast-control-member-based flood thresholds showed significant biases compared with the ERA5-based thresholds, confirming the benefit of using ensemble reforecasts.

4.6 | Practical recommendations for flood applications

Severe problems can arise in flood forecasting because of the potential issue with inconsistencies between the representation of extreme event frequencies in the thresholds and the forecasts, due to the biases that might be present especially for longer lead times. We recommend that forecast system developers should evaluate these potential inconsistencies for themselves using the methodology presented in this paper. We further recommend that this should be carried out with the use of reforecasts where they are available. But even where this is not the case, attempts should be made to diagnose the biases in the climatological data and the available historical forecasts for potential inconsistencies. Without addressing this inconsistency issue, the reliability and skill of the forecast flood events, and thus the quality of the flood warnings, could be substantially reduced, which could strongly impact on the decision-making process and ultimately lead to loss of confidence in the products.

In addition, even though the reanalysis-based flood thresholds are proven to be preferred in the first days of the forecast range, the difference in forecast skill to the reforecast-based thresholds is small. We recommend that it is both sensible and practical using the ensemble reforecasts for computing the flood thresholds for all forecast lead times and flood severity levels. Similarly, the best performing thresholds for the more impactful high floods (below 20% AEP) were generated from the median of a large number of random ensemble member selections from each reforecast. Although they are not necessarily the most favourable thresholds for smaller floods, they are the best overall choice and are recommended to be used for flood predictions across all flood severities and forecast lead times.

This study highlighted that flood forecasting applications, such as GloFAS, which use flood thresholds generated from a single time series (reanalysis or observation), can greatly benefit from using ensemble-reforecast-based thresholds instead, as a practical and effective way to resolve inconsistencies between forecasts and flood thresholds, and therefore increasing the flood forecast skill.

5 | CONCLUSIONS

Using reliable thresholds in global flood forecasting, that truly reflect the flood event frequencies of the real-time ensemble forecasts across all forecast lead times, is very important. The generation of flood signals with such thresholds can provide the highest forecast reliability and skill, which then gives the best chance to create trust in the users for the application.

In this paper, different annual maxima sampling methods were analysed to generate flood thresholds, using both GloFAS-ERA5 river discharge reanalysis and ensemble reforecasts. The flood thresholds were compared and their impact on the forecast reliability and skill was evaluated.

Reanalysis-based thresholds were found appropriate for the first 1–4 days of the 30-day (depending on the flood severity level) forecast range only. For longer lead times, both global average forecast reliability and skill deteriorate, effectively due to the increasing forecast biases over large parts of the world not accounted for in the reanalysis-based thresholds. The ensemble-reforecast-based thresholds provide increasing improvement over the reanalysis-based thresholds for up to the evaluated day 30 lead time. Additionally, using flood thresholds that sample the full ensemble in the reforecast, was

found to be overperforming a simple, single member sampling strategy (e.g., using the control reforecast), with generally better reliability and higher skill of the forecast probability.

The results of this study suggest that acknowledging the large uncertainty coming from the data sampling method in flood threshold generation is a crucial step in understanding and improving forecast skill, so that the system configuration that provides the highest reliability and lowest error globally can be found. In turn, better flood forecasts and better flood warnings could be delivered to the public, increasing the confidence and uptake of these products. Ultimately, the increase in confidence in the flood forecasts should result in better flood preparedness for humanitarian and civil protection partners, potentially reducing damages and casualties world-wide.

ACKNOWLEDGEMENTS

Ervin Zsoter's PhD is supported by the Wilkie Calvert Co-Supported PhD Studentships at the University of Reading. Ervin Zsoter and Christel Prudhomme were supported by the Copernicus Emergency Management Service—Early Warning Systems (CEMS-EWS [EFAS]). Hannah Cloke is supported by the TENDERLY project: Towards END-to End flood forecasting and a tool for ReaL-time catchment susceptibility UK NERC Flooding From Intense Rainfall (FFIR) programme, NE/K00896X/1. Elisabeth Stephens and Hannah Cloke are supported by the FATHUM project: Forecasts for AnTicipatory HUManitarian Action funded by UK NERC as part of their Science for Humanitarian Emergencies & Resilience (SHEAR) programme, NE/P000525/1.

AUTHORS' CONTRIBUTIONS

Ervin Zsoter designed the experiment, carried out the flood threshold analysis, and led the writing of the manuscript. Hannah Cloke and Liz Stephens assisted with posing the research question and designing the analysis. Christel Prudhomme and Florian Pappenberger helped designing the research methodology. All authors assisted with writing the manuscript.

DATA AVAILABILITY STATEMENT

The GloFAS-ERA5 river discharge reanalysis is openly available from the Copernicus Climate Change Service Climate Data Store. The GloFAS reforecasts will be made available through an ECMWF data repository in due course. The annual maxima time series and the related flood thresholds for all the analysed sampling methods are available upon request from the authors at ECMWF.

ORCID

Ervin Zsoter https://orcid.org/0000-0002-7998-0130

REFERENCES

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS—Global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, *17*, 1161–1175. https://doi.org/10.5194/hess-17-1161-2013

Alfieri, L., Berenguer, M., Knechtl, V., Liechti, K., Sempere-Torres, D., & Zappa, M. (2015). Flash flood forecasting based on rainfall thresholds. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. Cloke, & J. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-40457-3_49-1

Alfieri, L., Zsoter, E., Harrigan, S., Hirpa, F., Lavaysse, C., Prudhomme, C., & Salamon, P. (2019). Range-dependent thresholds for global flood early warning. *Journal of Hydrology*, *4*, 100034. https://doi.org/10.1016/j.hydroa.2019.100034

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., J.C.M. Andersson, J.C.M, Hasan A., & Pineda, L. (2020) Global catchment modelling using world-wide HYPE (WWH), open data and stepwise parameter estimation *Hydrology and Earth System Sciences*, *24*, 535–559, https://doi.org/10.5194/hess-24-535-2020.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., & Betts, A. K. (2009). A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, *10*, 623–643. https://doi.org/10.1175/2008JHM1068.1

Balsamo, G., Pappenberger, F., Dutra, E., Viterbo, P., & van den Hurk, B. (2011). A revised land hydrology in the ECMWF model: A step towards daily water flux prediction in a fully-closed water cycle. *Hydrological Processes*, *25*, 1046–1054. https://doi.org/10.1002/hyp.7808

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., … Cloke, H. L. (2016). Continental and global scale flood forecasting systems. *WIREs Water*, *3*, 391–418. https://doi.org/10.1002/wat2.1137

Emerton, R., Cloke, H., & Stephens, E. (2017). Complex picture for likelihood of ENSO-driven flood hazard. *Nature Communications*, *8*, 14796. https://doi.org/10.1038/ncomms14796

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., … Pappenberger, F. (2018). Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS—seasonal v1.0. *Geoscientific Model Development*, *11*, 3327–3346. https://doi.org/10.5194/gmd-11-3327-2018

Gumbel, E. J. (1941). The return period of flood flows. *Annals of Mathematical Statistics*, *12*(2), 163–190.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., … Zuo, H. (2018) *Operational global reanalysis: progress, future directions and synergies with NWP*. ERA Report, ECMWF, UK. https://doi:10.21957/tkic6g3wm

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., … Thepaut, J-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 1–51. https://doi.org/10.1002/qj.3803

Hirpa, F. A., Salamon, P., Alfieri, L., Pozo, J. T., Zsoter, E., & Pappenberger, F. (2016). The effect of reference climatology on

global flood forecasting. *Journal of Hydrometeorology*, *17*, 1131–1145. https://doi.org/10.1175/JHM-D-15-0044.1

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., & Dadson, S. J. (2018). Calibration of the global flood awareness system (GloFAS) using daily streamflow data. *Journal of Hydrology*, *566*, 595–606. https://doi.org/10.1016/j.jhydrol.2018.09.052

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B: Methodological*, *52*, 105–124.

Hsu, W. R., & Murphy, A. H. (1986). The attributes diagram: A geometrical framework for assessing the quality of probability forecast. *International Journal of Forecasting*, *2*, 285–293.

Kjeldsen, T. R., Lamb, R., & Blazkova, S. D. (2014). Uncertainty in flood frequency analysis. In K. Beven & J. Hall (Eds.), *Applied uncertainty analysis for flood frequency analysis*. London: Imperial Collage Press.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., … Smolarkiewicz, P. K. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, *143*, 2315–2339. https://doi.org/10.1002/qj.3094

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*, 187–201. https://doi.org/10.1029/2012WR012557

Stephens, E., Day, J. J., Pappenberger, F., & Cloke, H. (2015). Precipitation and floodiness. *Geophysical Research Letters*, *42*, 10316–10323. https://doi.org/10.1002/2015GL066779

Thielen, J., Bartholmes, J., Ramos, M.-H., & de Roo, A. (2009). The European flood alert system—Part 1: Concept and development. *Hydrology and Earth System Sciences*, *13*, 125–140. https://doi.org/10.5194/hess-13-125-2009

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., … Stephens, E. M. (2019). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrology and Earth System Sciences*, *23*(7), 3057–3080. https://doi.org/10.5194/hess-23-3057-2019

Tsonevsky, I., Doswell, C. A., & Brooks, H. E. (2018). Early warnings of severe convection using the ECMWF extreme forecast index. *Weather and Forecasting*, *33*, 857–871. https://doi.org/10.1175/WAF-D-18-0030.1

van der Knijff, J. M., Younis, J., & de Roo, A. P. J. (2010). LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, *24*, 189–212. https://doi.org/10.1080/13658810802549154

Verkade, J. S., Brown, J. D., Reggiani, P., & Weerts, A. H. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, *501*, 73–91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast scores. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1889–1899. https://doi.org/10.1002/qj.2256

Wentao, L., Qingyun, D., Chiyuan, M., Aizhong, Y., Wei, G., & Zhenhua, D. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, *4*, e1246. https://doi.org/10.1002/wat2.1246

WMO Guidelines on the calculation of Climate Normals, WMO-No. 1203 (2017).

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., & Robertson, D. E. (2020). Ensemble flood forecasting: Current status and future opportunities. *WIREs Water*, *7*(3), e1432. https://doi.org/10.1002/wat2.1432

Yuan, X., Wood, E. F., & F, E. (2012). Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research*, *48*, W12519. https://doi.org/10.1029/2012WR012256

Zsoter, E., Pappenberger, F., & Richardson, D. (2014). Sensitivity of model climate to sampling configurations and the impact on the extreme forecast index. *Meteorological Applications*, *22*, 236–247. https://doi.org/10.1002/met.1447

Zsoter, E., Pappenberger, F., Smith, P., Emerton, R., Dutra, E., Wetterhall, F., … Balsamo, G. (2016). Building a multi-model flood prediction system with the TIGGE archive. *Journal of Hydrometeorology*, *17*, 2923–2940. https://doi.org/10.1175/JHM-D-15-0130.1

Zsoter, E., Cloke, H., Stephens, E., de Rosnay, P., Muñoz-Sabater, J., Prudhomme, C., & Pappenberger, F. (2019). How well do operational numerical weather prediction setups represent hydrology. *Journal of Hydrometeorology*, *14*, 1533–1552. https://doi.org/10.1175/JHM-D-18-0086.1

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Zsoter E, Prudhomme C, Stephens E, Pappenberger F, Cloke H. Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting. *J Flood Risk Management*. 2020;13:e12658. https://doi.org/10.1111/jfr3.12658