

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

GloFAS Map Viewer: <http://www.globalfloods.eu/>

Ervin Zsoter^{1,2}, Christel Prudhomme¹, Florian Pappenberger¹, Elisabeth Stephens², Hannah Cloke^{2,3}

(1) European Centre for Medium-Range Weather Forecasts, ECMWF, UK

(2) Department of Geography and Environmental Science, University of Reading, UK

(3) Department of Earth Sciences, Uppsala University, Uppsala, Sweden

ervin.zsoter@ecmwf.int

Zsoter, E., Prudhomme, C., Stephens, E., Pappenberger, F. and Cloke, H.: Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting, J. Flood Risk Manag., doi:10.1111/jfr3.12658, 2020.



European
Commission

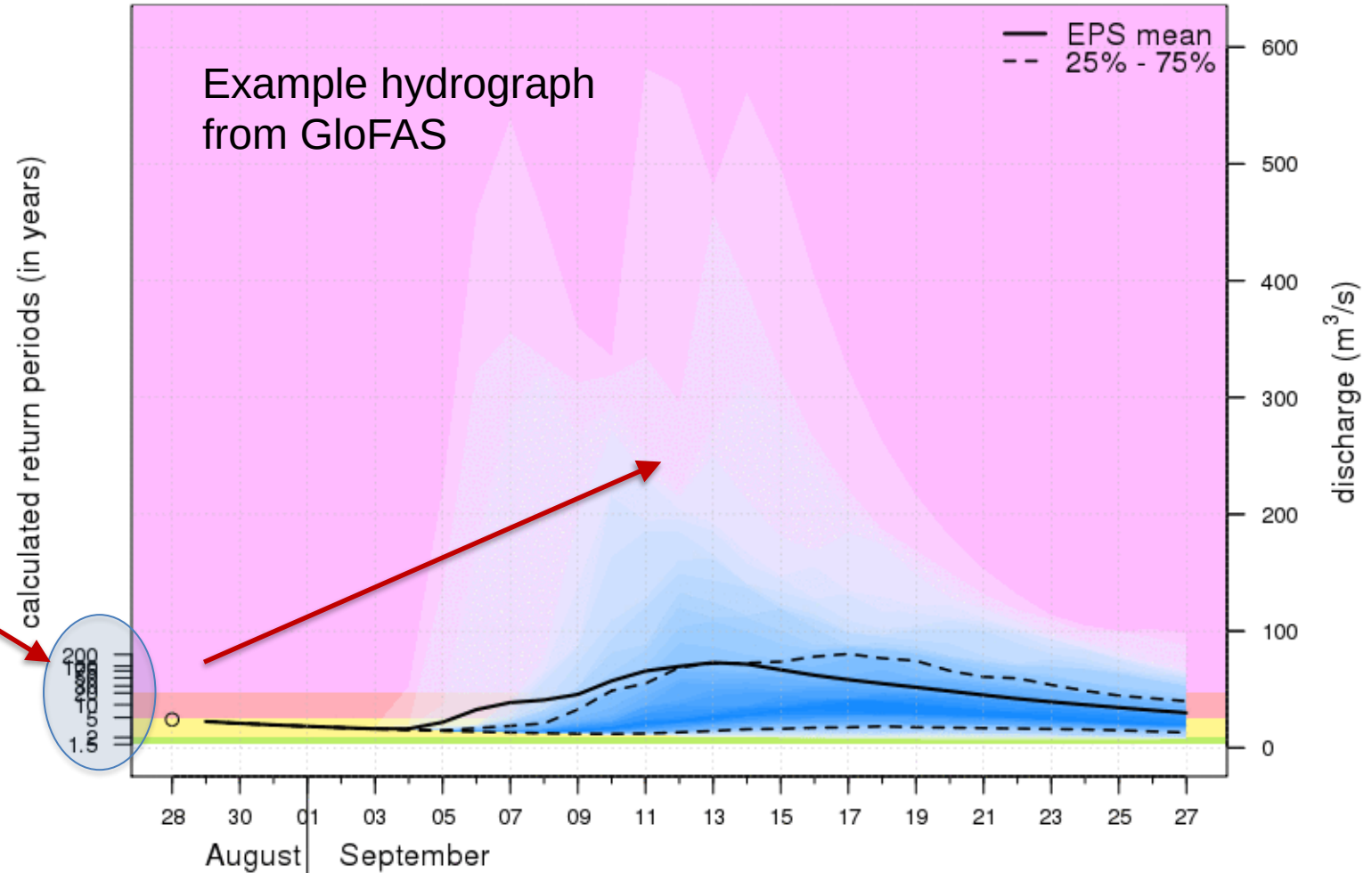


University of
Reading



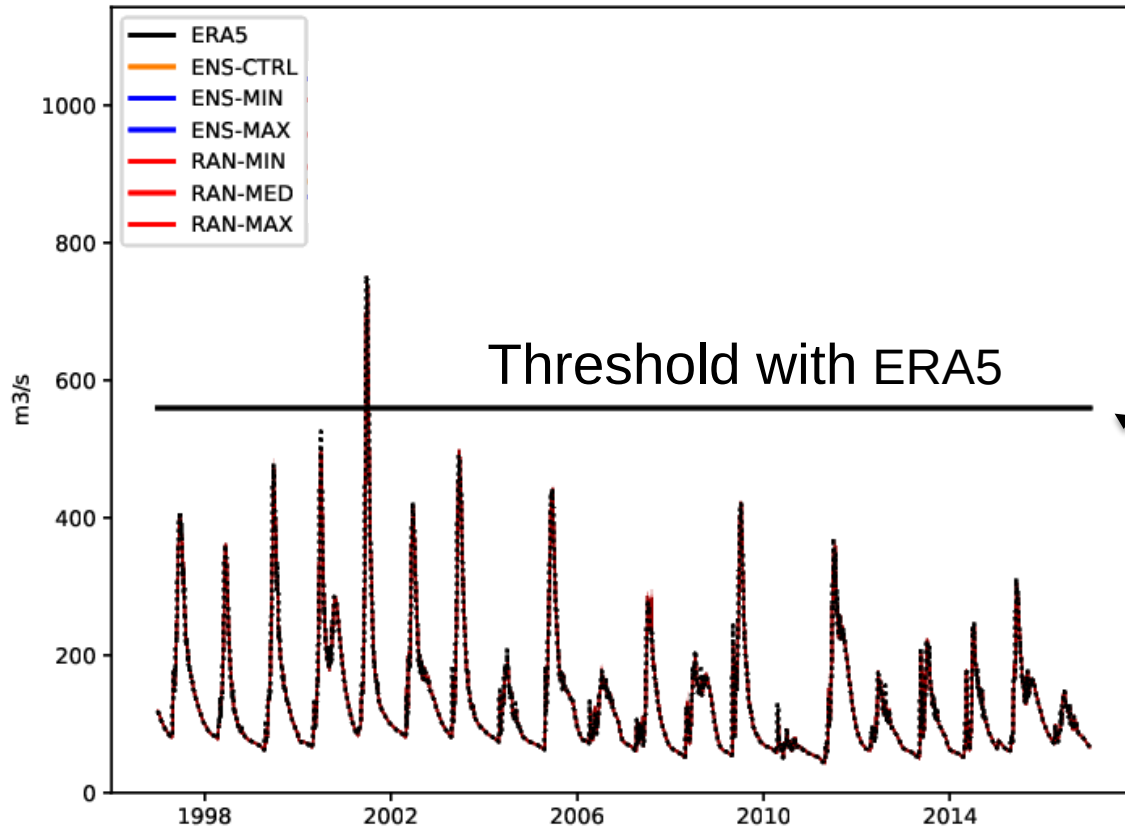


- Flood signals in global flood forecasting systems (such as the Global Flood Awareness System – GloFAS) are derived by comparing real-time forecasts to flood thresholds
- Flood thresholds are often derived from reanalysis and are fixed for all forecast lead times (30-day horizon in GloFAS)
- Occasionally very high return periods predicted
- Are these realistic or just an artefact of the changing extreme event behaviour in the forecast across lead times?
- Complex forecast biases to blame?!





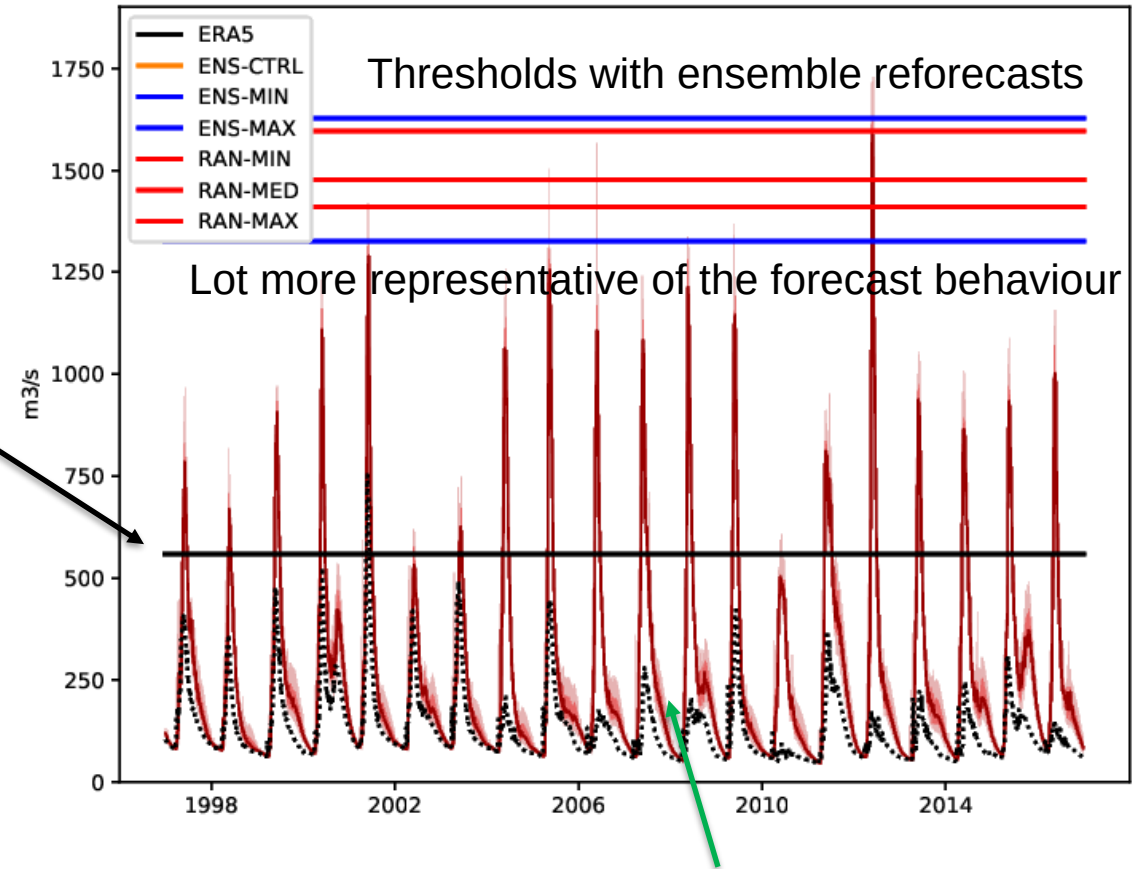
Day-1



1997-2016 reforecasts
Initialised from ERA5-simulation

ERA5: black dashed line
Ensemble reforecasts: red lines

Day-30



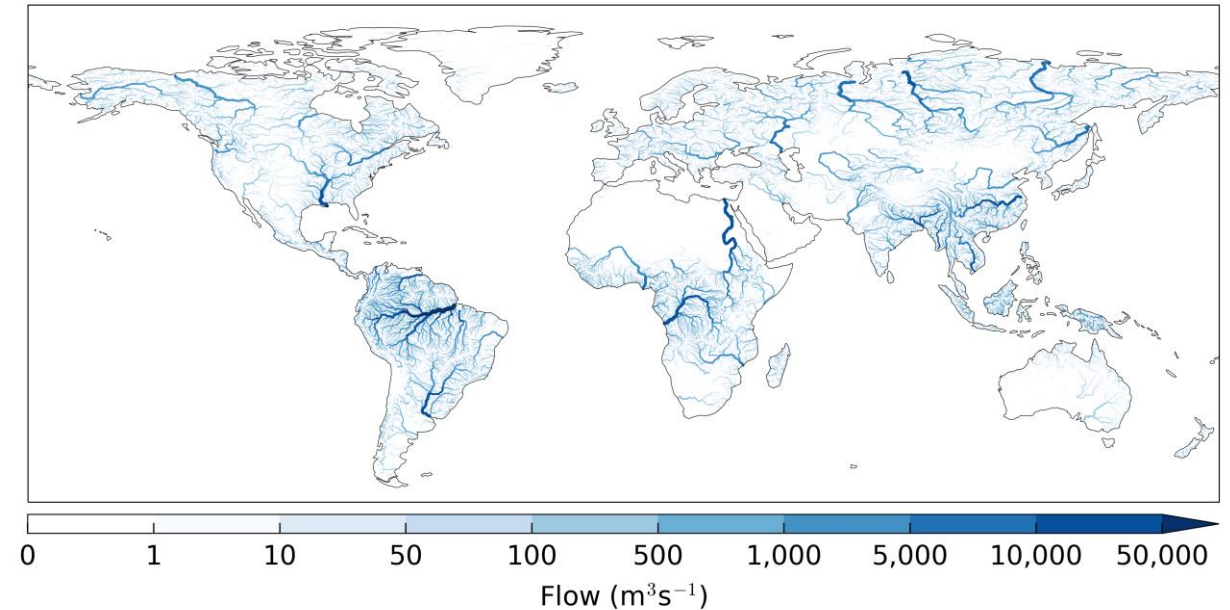
Ensemble reforecasts are
systematically higher at longer
lead times



GloFAS-ERA5 reanalysis

- River discharge reanalysis (Harrigan et al., 2020)
- GloFAS-ERA5 uses the high-resolution ERA5 component at 31 km horizontal resolution
- Produced by coupling surface and sub-surface runoff from ERA5 (output of the HTESSEL land surface model) with the LISFLOOD hydrological and channel routing model.
- Global gridded dataset with a horizontal resolution of 0.1° at a daily time step
- Openly available in near real time from the Copernicus Climate Change Service Climate Data Store

Mean daily river discharge from 1979-2018 for GloFASv2.1 reanalysis



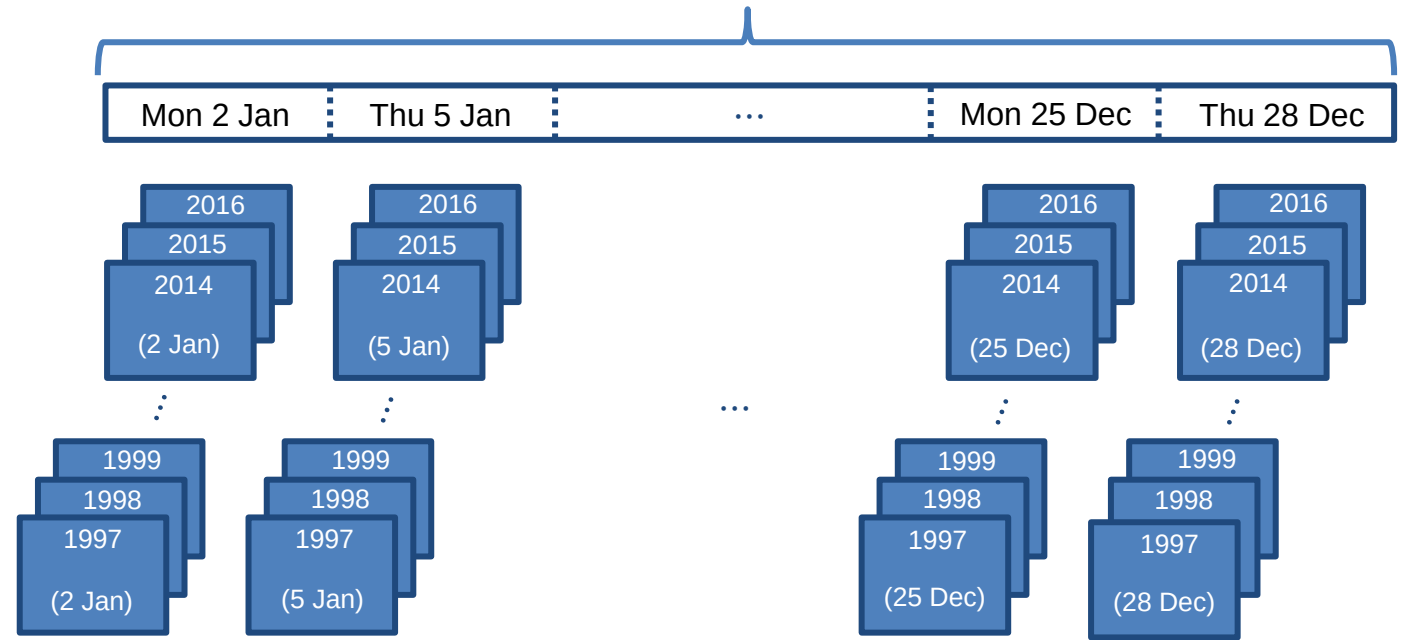
Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, *Earth Syst. Sci. Data*, 12, 2043–2060, <https://doi.org/10.5194/essd-12-2043-2020>, 2020.



GloFAS ensemble reforecasts

- Ensemble river discharge reforecasts generated for past dates (20 years) by the same GloFAS system that is used for real time forecasts
- Produced twice weekly (Mondays and Thursdays) with 11-members and 30-day lead time
- Forced by ECMWF meteorological ensemble reforecasts
- Initialised from GloFAS-ERA5
- In this study, reforecasts produced for 2017 were used (for each year in 1997–2016 with 2017-Monday and 2017-Thursday initial times).
- Altogether 2080 reforecasts (104 in each year)

GloFAS reforecasts for year 2017

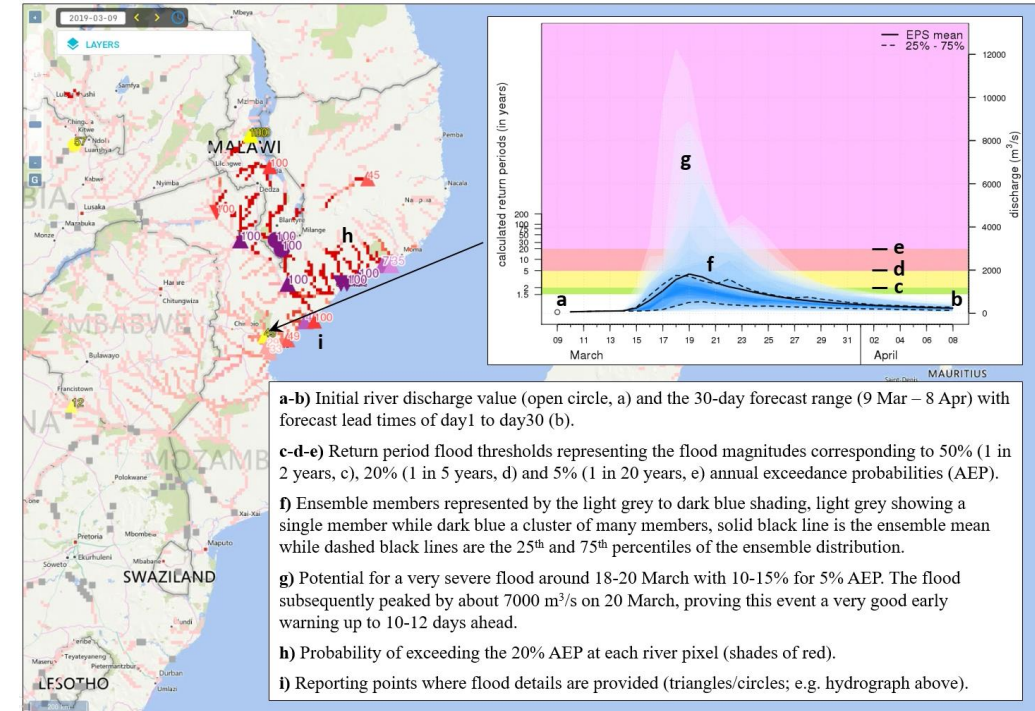


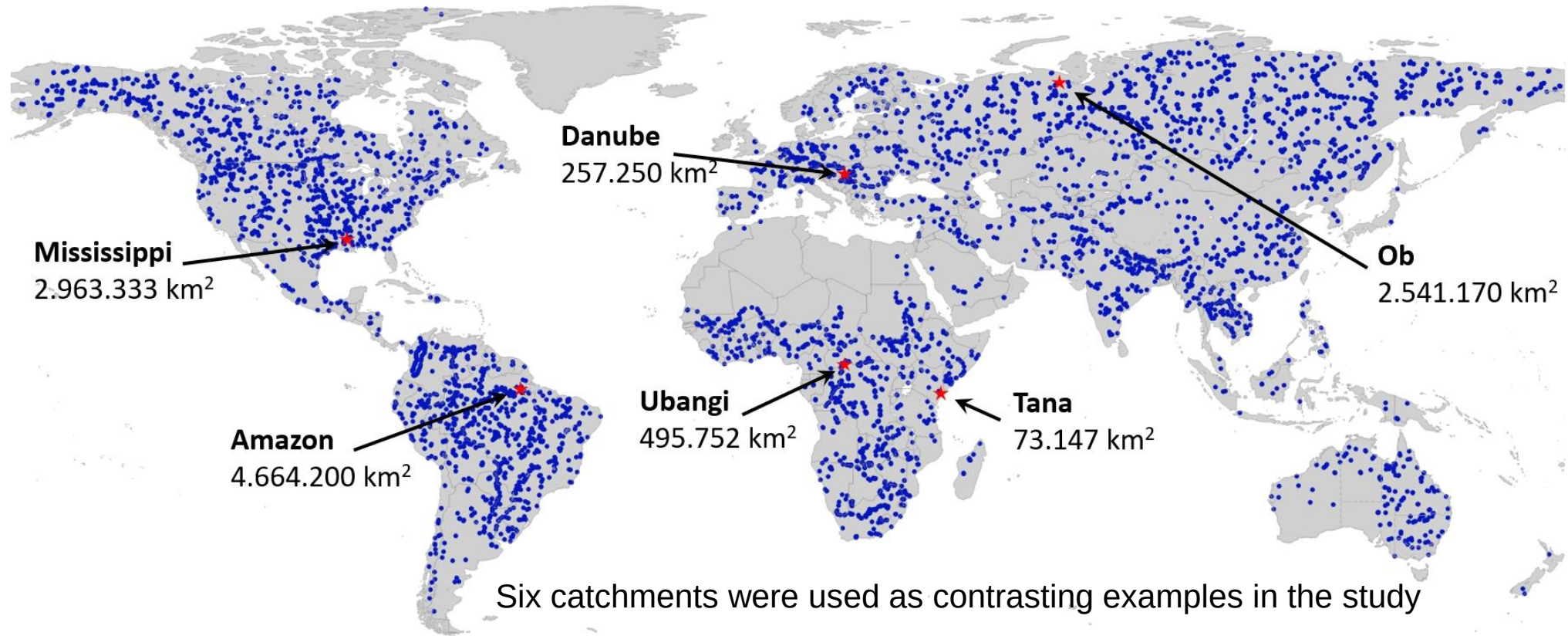
Harrigan, S., Zoster, E., Cloke, H., Salamon, P., and Prudhomme, C.: Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System, Hydrol. Earth Syst. Sci. Discuss. [preprint], <https://doi.org/10.5194/hess-2020-532>, in review, 2020.



Flood thresholds

- In GloFAS, flood quantiles of three severity levels (2-, 5-, and 20-year return periods) are used as flood thresholds
- A return period T is an estimate of the likelihood of an event to occur, 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 $AEP = 100/T$)
- In this study, 4 severity levels were analysed, 50, 20, 10 and 5% AEP
- The flood thresholds were computed (as in GloFAS) by fitting the Gumbel Extreme Value Distribution (EVD) to the annual maximum river discharge sample using the method of L-Moments
- This method is appropriate for relatively small sample sizes (20 values for the 20 reforecast years of 1997 to 2016)





- The study was based on the GloFAS diagnostic station network (Figure 2)
- This network provides a global coverage and includes all points where daily historical river discharge observations are available for the GloFAS team
- Catchments that have 50% AEP magnitude below 20 m³/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



- Question no.1: How adequate it is to use a reanalysis dataset to define flood thresholds and apply them for all forecast lead times?
- Question no.2: How best to use reforecast ensemble information in the flood threshold generation to improve flood forecast performance?
- The impacts of the choice of data source (reanalysis or reforecasts) and of the annual maximum sampling strategies (from the reforecasts) were analysed by comparing the flood threshold magnitudes (for each catchment) and the resulting forecast reliability and skill benefits for four different flood severity levels
- The flood forecast performance was evaluated for day 1 to day 30 lead times on the 2080-forecast sample 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 and the reliability diagram. This step assesses both the skill and reliability in the resulting probability forecasts of exceeding the thresholds

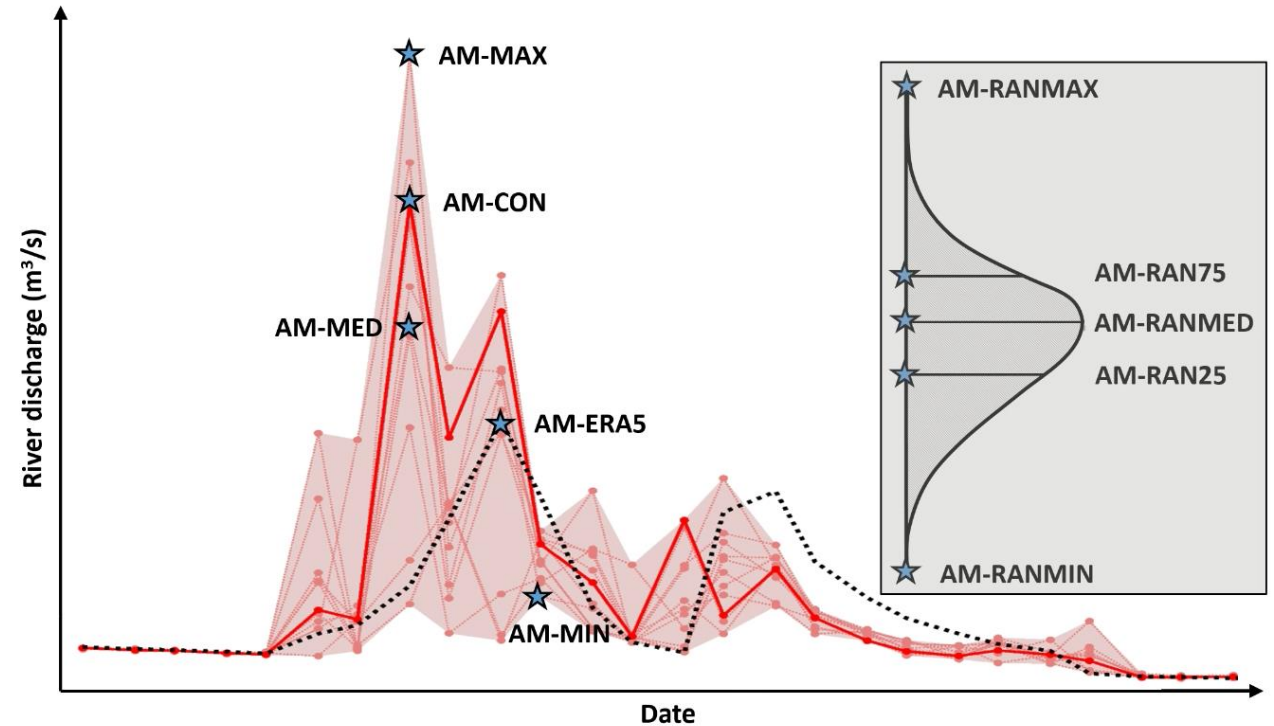


Different sampling strategies for annual maxima

- Benchmark set: GloFAS-ERA5 reanalysis (independent from the lead time)
- Reforecast set: control member, plus the minimum, median and maximum values from each run of the 11-member GloFAS reforecasts
- Extended reforecast set: 1,000 times randomly selecting from the 11 ensemble members from each GloFAS reforecast

This resulted in 1 + 4 + 1,000 threshold values (for each severity level, lead time and catchment)

From the 1,000 random member-based thresholds the 25-th and 75-th percentiles, minimum, median and maximum values were analysed further

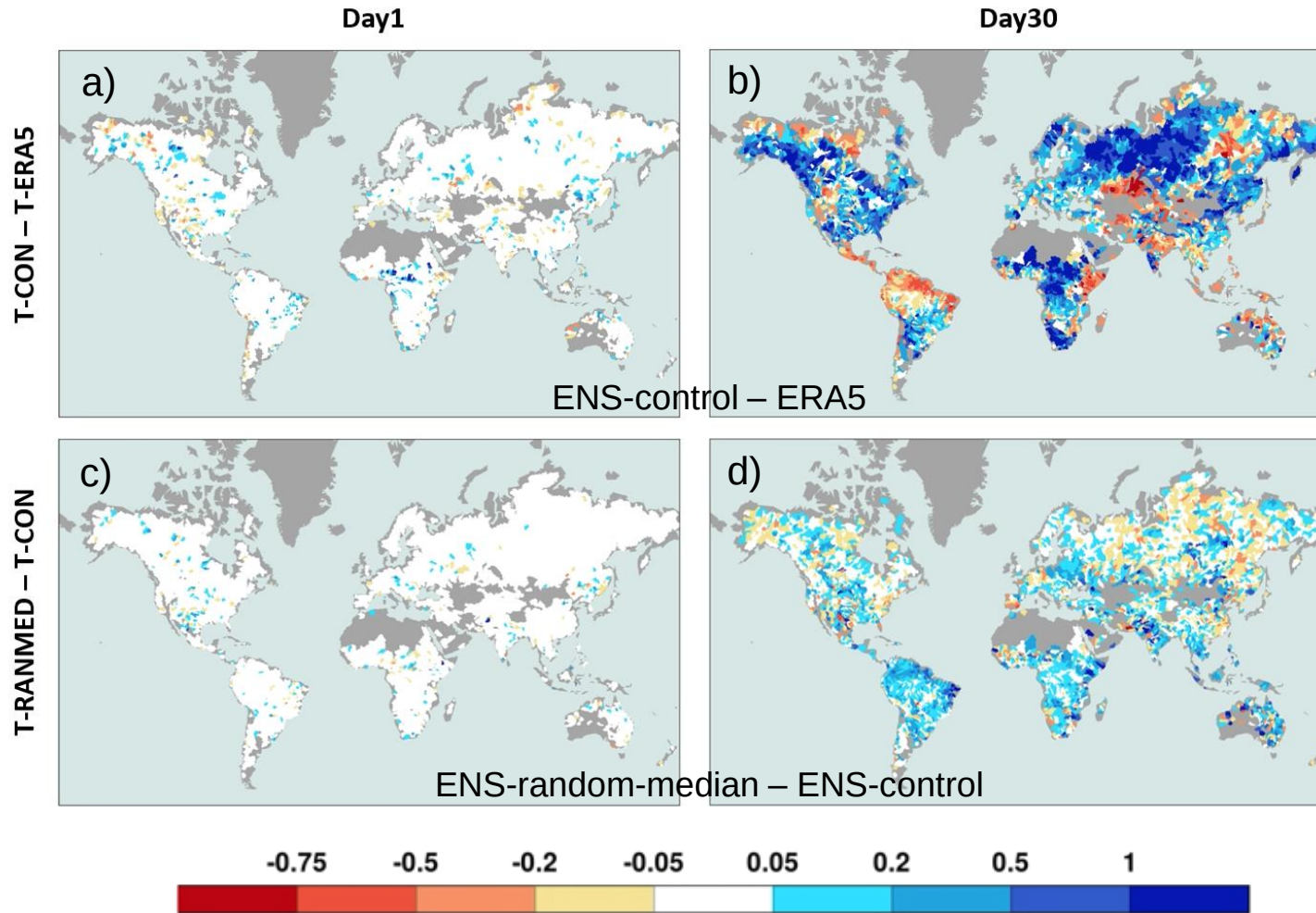


Version	Time series set	Version	Time series set
T-ERA5	Benchmark set	T-RANMIN	Minimum extended reforecast set
T-CON	Control member reforecast set	T-RAN25	25th percentile extended reforecast set
T-MIN	Minimum member reforecast set	T-RANMED	Median extended reforecast set
T-MED	Median member reforecast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



How similar are the flood thresholds?

%-difference between thresholds for 10% AEP

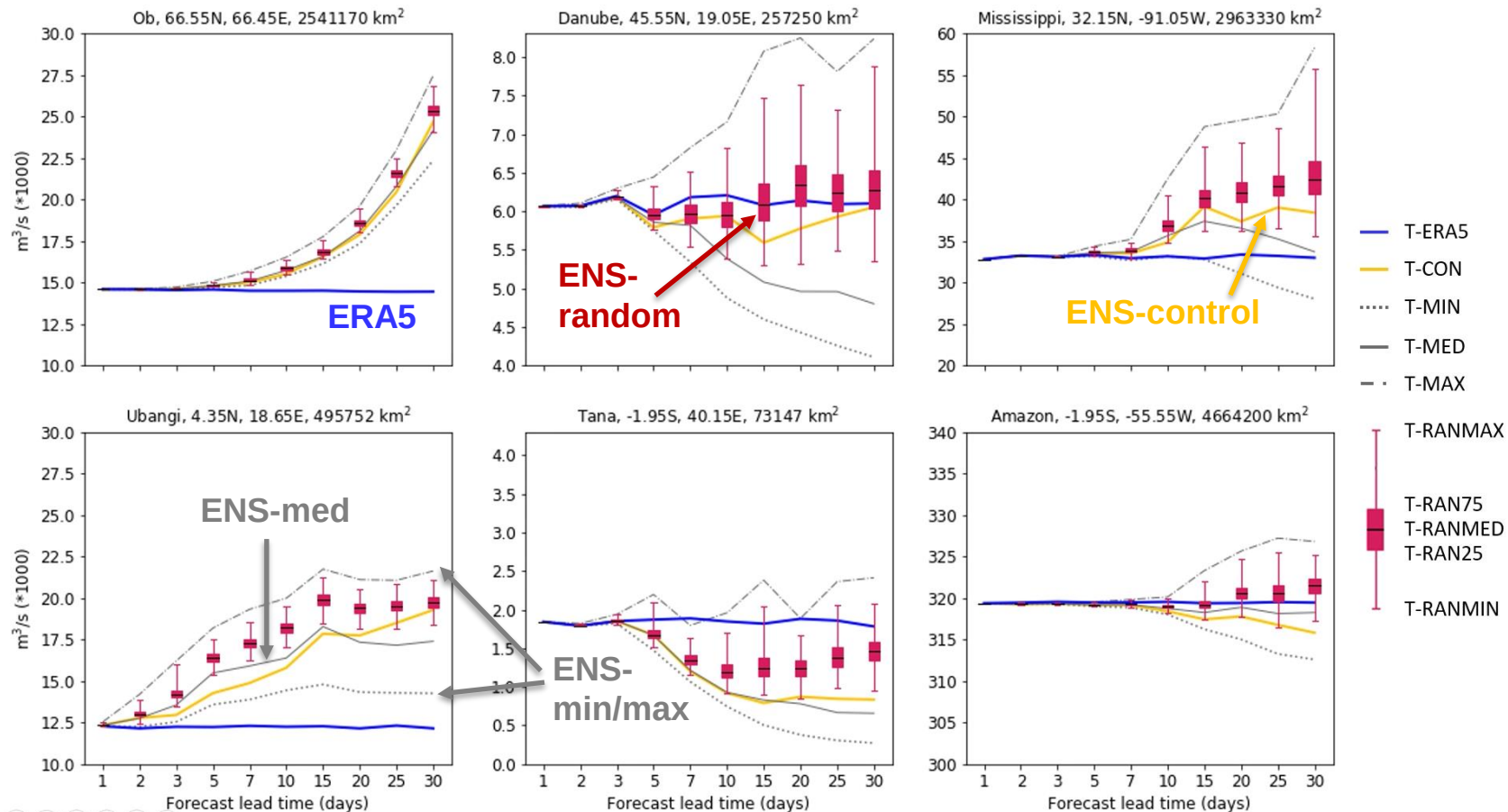


- At day 1, all sampling strategies produce similar thresholds
- However, by day 30, T-ERA5 becomes very different to the reforecast-based thresholds (over 50% difference for large parts of the world) with large variability geographically
- Using only the ensemble control or the full reforecast ensemble, does also create noticeable differences by day 30 (although smaller differences than between T-ERA5 and the reforecast-based thresholds)



How similar are the flood thresholds?

10% AEP flood thresholds as a function of lead time for six contrasting catchments (see catchment locations above)

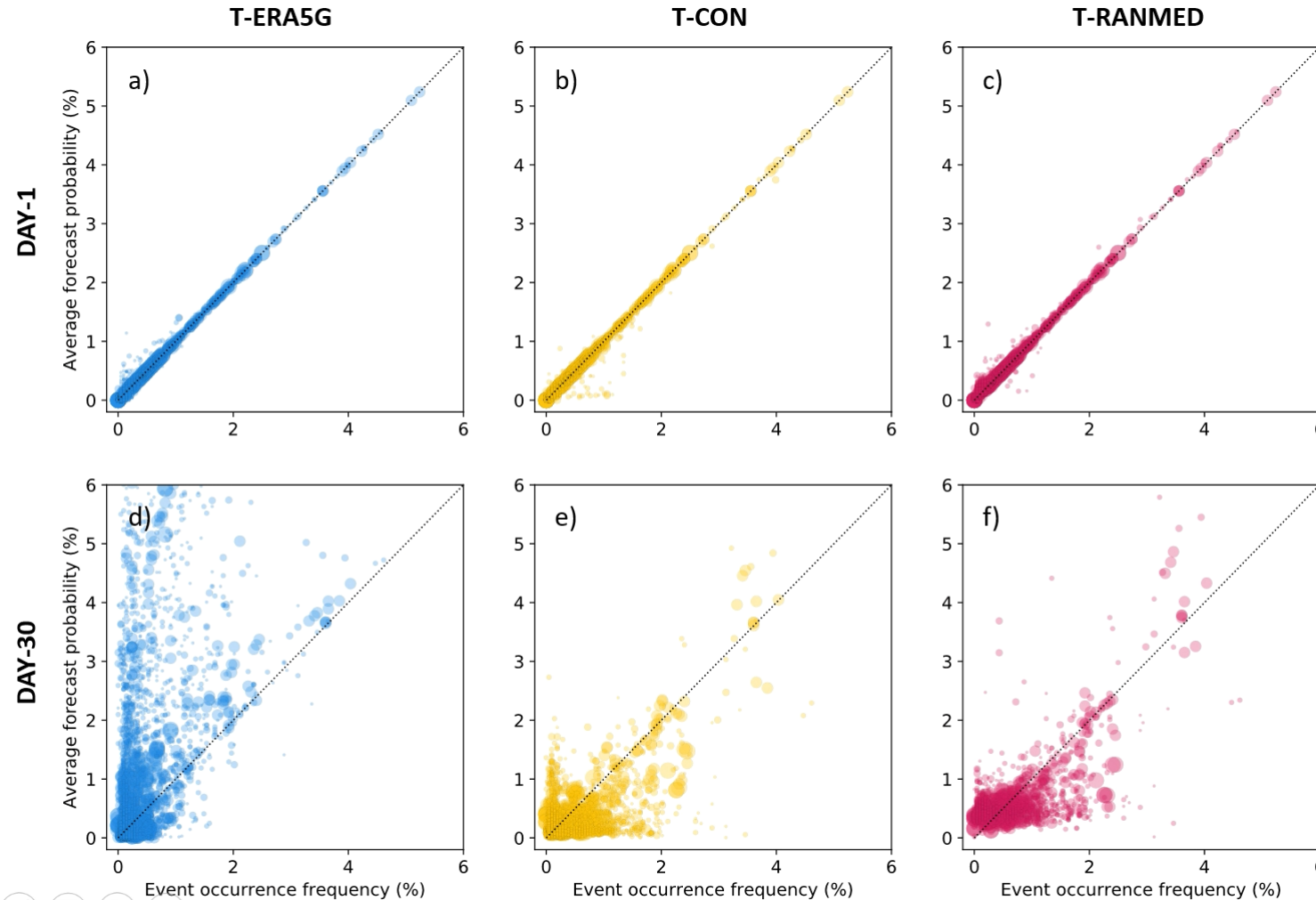


- The influence of the ensemble reforecast sampling strategy gets larger with the increasing forecast lead time
- Uncertainty in the extended reforecast set (1000 random thresholds) varies from catchment to catchment, and can be very large (such as river Danube) or much smaller (such as river Ob)
- The control-member-based flood thresholds are dominantly below the envelope of the ensemble reforecast (see also the dominantly negative values in figure (d) on the previous slide)



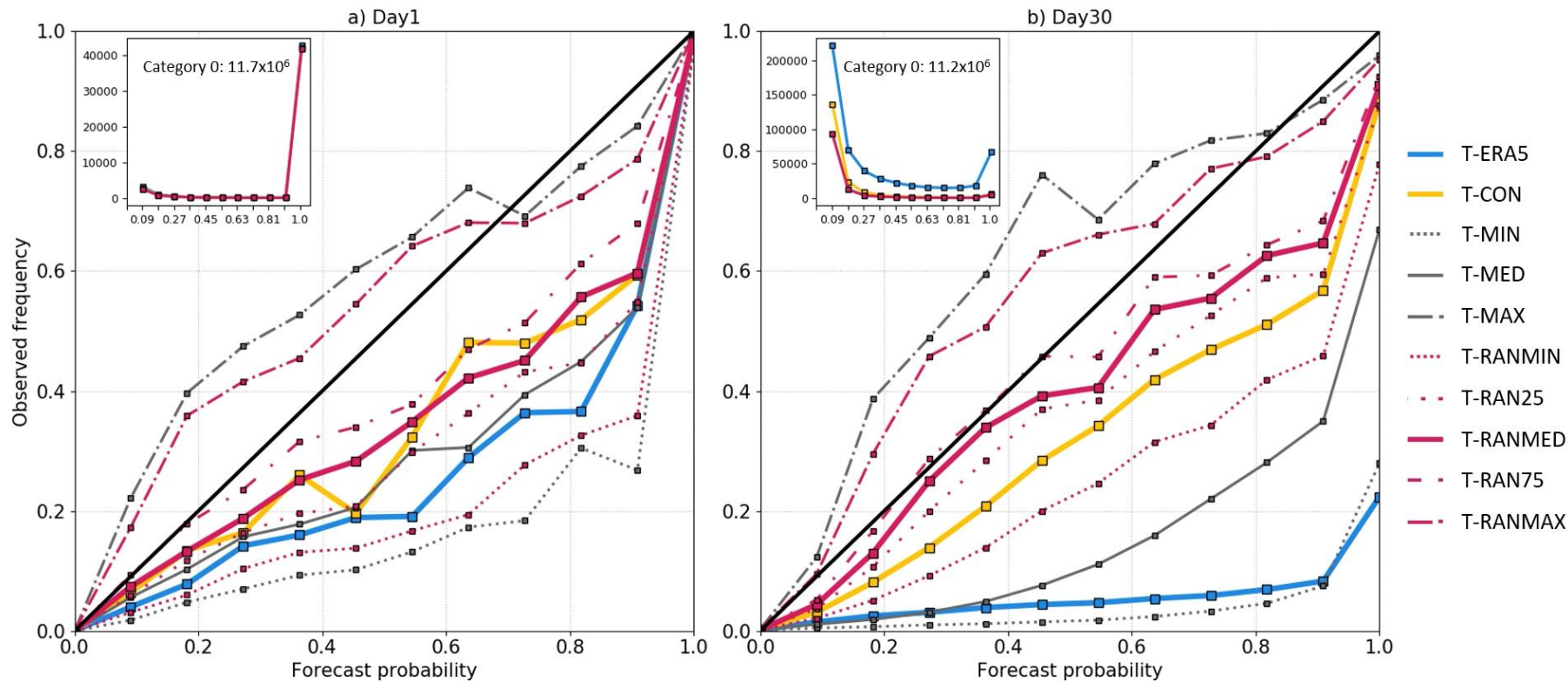
How reliable are the forecast probabilities

10% AEP



Flood forecast probabilities $\stackrel{?}{\Leftrightarrow}$ Flood occurrence frequencies

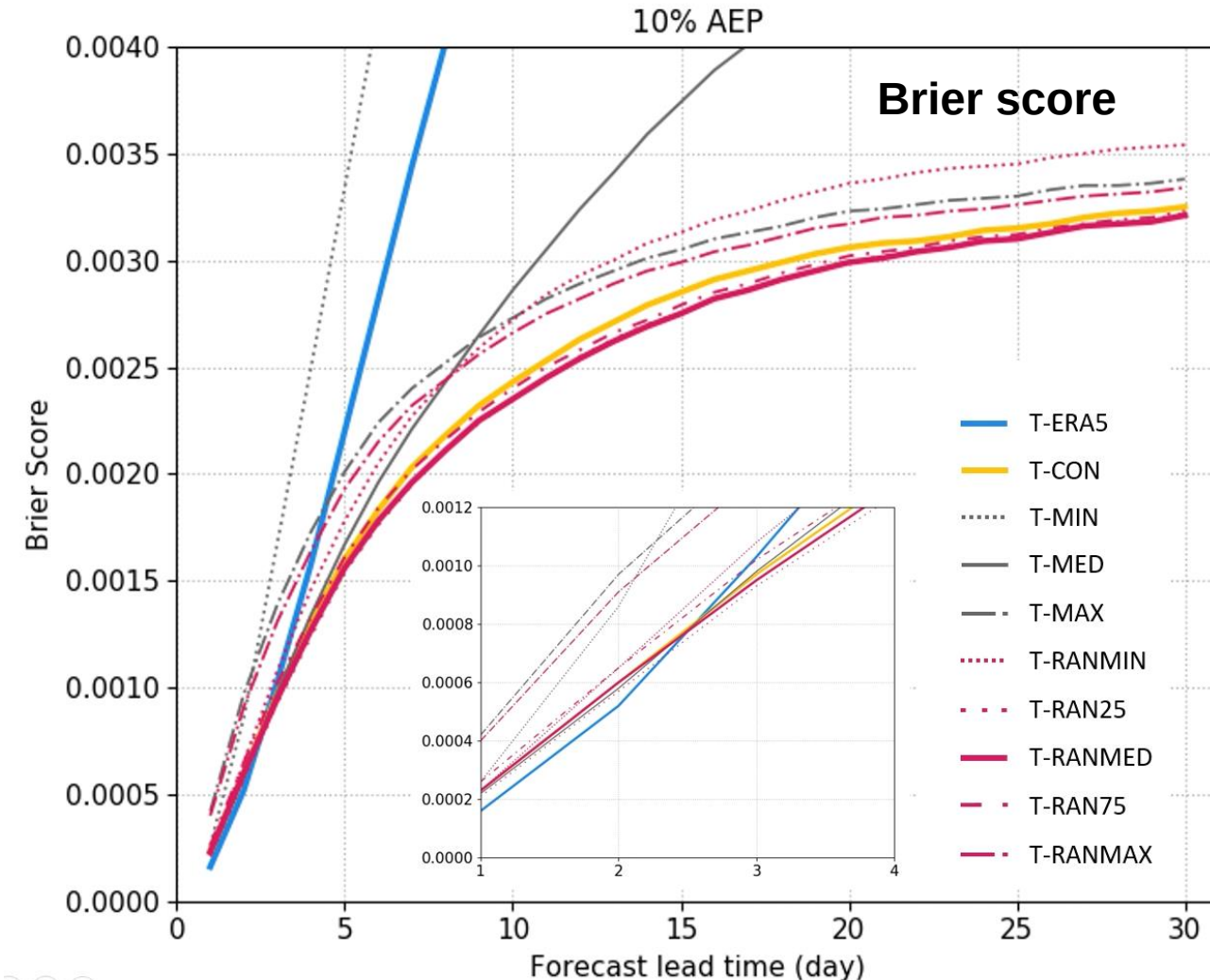
- Number of flood events occurring in the forecast depends on the flood thresholds
- At day 1 the match is near perfect
- By day 30 the match is lost, with the benchmark T-ERA5 showing the largest mismatch (probabilities are too high)
- Reforecasts-based thresholds show a clear improvement
- Results based on T-RANMED are slightly better than with T-CON (better clustering around the diagonal line)



- Reliability of the flood forecasts with the T-ERA5 (fixed reanalysis-based) thresholds is low, especially for day longer lead times
- The reforecast-based flood thresholds can greatly improve the flood forecast reliability (only exceptions are T-MIN and T-RANMIN for short lead time)
- T-MAX and T-RANMAX tend to systematically underestimate flood event frequency up to 70–80% forecast probability, whilst overestimation of flood event numbers 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



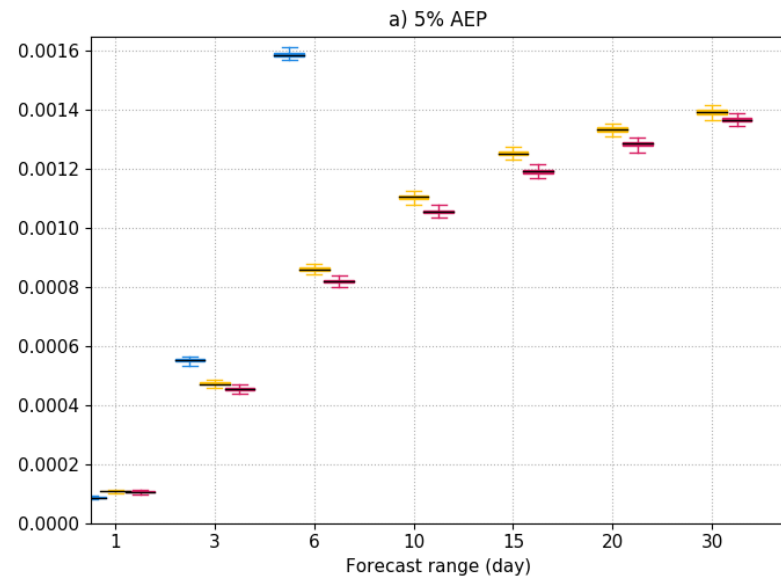
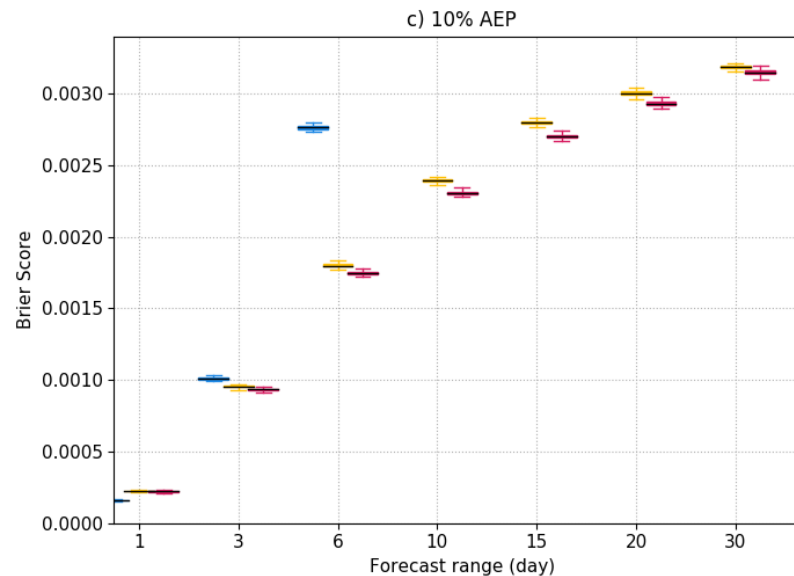
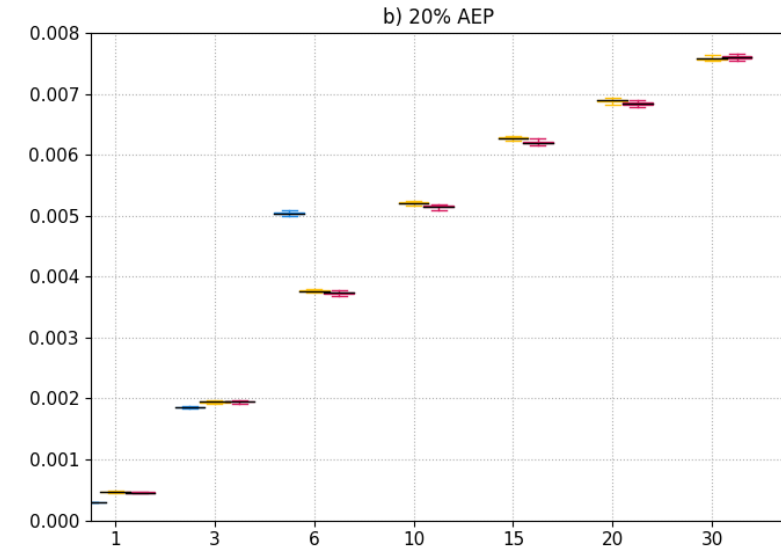
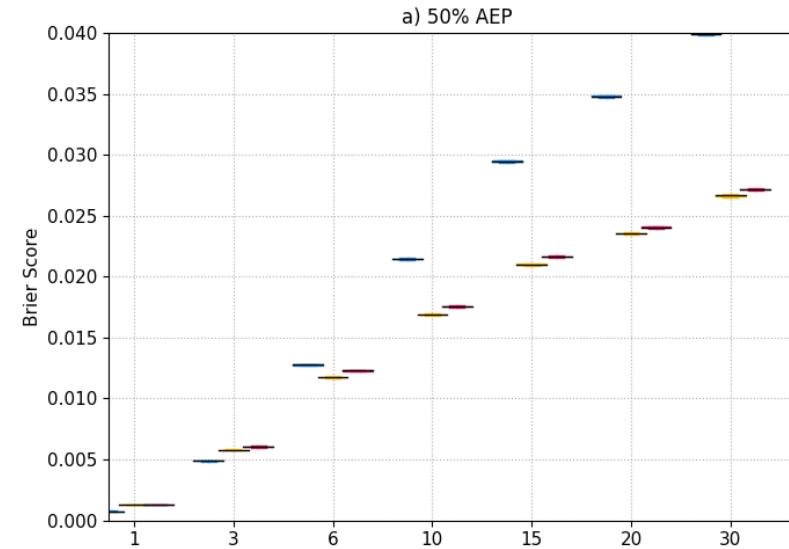
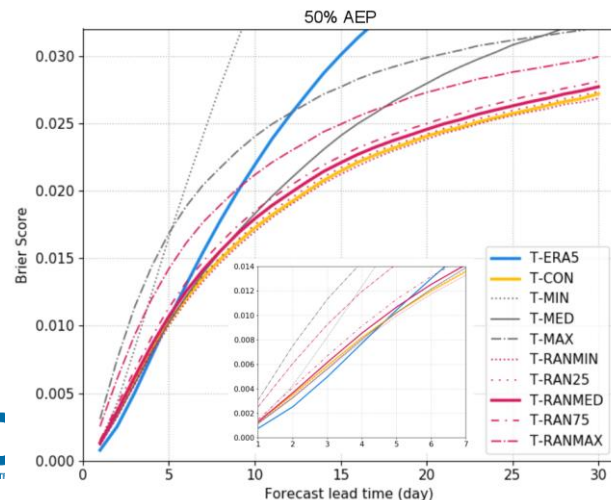
- The benchmark T-ERA5 threshold is best only up to day 2 lead time for 10% AEP (blue line below the other lines). For other flood severity levels this varies from day 1 (5% AEP) to day 4 (50% AEP)
- From day 3, the flood forecasts with reforecast-based thresholds become more and more skillful than with T-ERA5 (the only exception is T-MIN)
- Both T-ERA5 and T-MIN (and somewhat also T-MED) show poor skill with multiple times higher Brier score values by day 30
- The best performance is achieved by the reforecast-based random median (T-RANMED) and the interquartile range boundaries (T-RAN25 and T-RAN75) and the slightly less skillful reforecast control threshold (T-CON), with skill slowly degrading with lead time





Impact on forecast reliability and skill

- The skill improvement, using T-RANMED over T-CON, is statistically significant at the 99% level from day 5–6 (tested by bootstrapping the dates in the verification sample)
- The pattern is very 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





Ensemble member independence

- Using ensemble members allows a better representation of the extreme events in the forecast climatology by increasing the sample size
- Although the members are correlated especially at short lead times, only one member was chosen from all twice weekly reforecasts, which made them an effectively independent realisations of the true underlying model climate distribution

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 potentially grow large, affecting the extreme event representation and the use of flood thresholds in medium to extended range hydrological forecast systems like GloFAS
- These biases can originate from the meteorological forcing and impact the hydrological simulations (mainly through precipitation)
- Another likely source for the biases is the land data assimilation impact which can open the water balance (e.g through adding or removing snow or soil moisture, such as in GloFAS with the HTESSEL land-surface model) and thus introduce biases which will also translate into the forecasts



Forecast post-processing

- Post-processing of the forecasts against the reference dataset, used to derive the flood thresholds (i.e. reanalysis), is a possible alternative to ensemble reforecast-based thresholds
- By removing biases in the forecasts (e.g., linear regression or quantile mapping), the extreme event representation of the forecasts would be expected to become similar to the behaviour of the reference dataset

Modelling system independence

- Although this study was done with GloFAS (using HTESSEL land-surface model), the main findings are expected to be independent of the modelling system, extreme value fitting method or sampling period length

Practical recommendations for flood applications

- Evolving biases in the forecast for different lead times can cause inconsistencies between extreme event representation in the thresholds and the forecasts
- We recommend that forecast system developers should evaluate these potential inconsistencies, using reforecasts where available
- This study highlighted that ensemble-reforecast-based thresholds are a practical and effective way to resolve inconsistencies between forecasts and flood thresholds
- This can greatly increase the forecast skill in flood forecasting applications impacted by biases, such as GloFAS, where the thresholds are generated from a single time series (reanalysis or observation)