**Peter Salamon Referee #2**

**General comment:**

The manuscript presents an approach to increase the sample size for the estimation of the frequency of flood events. The approach is based on pooling of reforecast ensemble members and has not been previously assessed for flood frequency analysis.

The paper is overall well written and structured. The approach presented in this manuscript is of high interest as estimating flood frequency in practice is often hampered by short observational records. The discussion section outlines the possible limitations of the approach. My main concern is related to the data used for the study. The study uses EFAS v3.0 historical simulations to assess whether the selected stations have a good performance when comparing simulations and observations. However, the EFAS reforecast data set used for the ensemble pooling is based on EFAS v4.0 which includes a completely new model calibration, upgrades to static fields for the hydrological model LISFLOOD and a change from a daily timestep to a 6 hourly timestep. Overall, EFAS model performance from v3.0 to v4.0 has increased significantly and therefore it is not recommended to select stations based on v3.0 and perform an analysis using reforecasts that are based on EFAS 4.0. As this has an impact on all results and analysis in the manuscript a major revision is required.

**Main concern:**

As described in the general comment EFAS reforecasts are based on EFAS v4.0 as is also indicated in the metadata on the Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/efas-reforecast?tab=overview ). However, EFAS historical simulations v3.0 were used to pre-select stations with a good fit between simulated and observed discharge. Given that EFAS 4.0 contains a completely new model calibration with more calibration stations (1137 for v4.0 instead of 717 stations for previous EFAS versions), and upgrades to static fields for the hydrological model LISFLOOD and a change from a daily timestep to a 6 hourly timestep as is described in detail in the EFAS wiki (see here: https://confluence.ecmwf.int/display/COPSRV/EFAS+v4.0 ) it is not recommendable to use EFAS v3.0 model performance to pre-select stations and then use those pre-selected stations with EFAS re-forecasts from EFAS v4.0.

**Reply:** *Thank you for highlighting that there were inconsistencies between the description of the calibration procedure we provided (which implied we were using EFAS 3.0) and the actual use of EFAS 4.0 (which we employed for both our station selection and analysis). In the analysis, we did compare EFAS 4.0 historical runs with observed GRDC data for model evaluation, i.e. the EFAS version used was consistent. However, our initial description of the calibration procedure suggested that we were using runs from EFAS 3.0 to select pre-stations. This is not the case and we updated the description of the calibration procedure to match the calibration procedure used in EFAS 4.0 as documented on the EFAS Wiki pages.*

Furthermore, the authors do not describe in detail how the simulated data was extracted from the EFAS simulations. EFAS output has a spatial resolution of 5km x 5km. The coarse spatial resolution of the hydrological model LISFLOOD used in EFAS requires an upscaling of the river drainage network from a high resolution datasest to the 5km x 5km grid scale. This means that coordinates of gauging stations cannot be used directly to extract simulated timeseries of discharge from the EFAS simulations as original gauging station coordinates may be located just next to the main river channel on the coarse grid scale. Instead, before extracting simulated time series it has to be checked whether the drainage area of the EFAS grid pixel corresponds to the drainage area as provided by the data provider (here GRDC). While smaller differences in the drainage area are expected due to the different spatial scales, if there is a large difference, it means that coordinates have to be shifted to ensure an adequate match. For this purpose the drainage area of the LISFLOOD/EFAS network is available on the C3S CDS (https://cds.climate.copernicus.eu/cdsapp#!/dataset/efas-historical?tab=overview ). This is especially important for gauging stations with very small drainage areas which seem to have been used predominantly in this study (Fig.

9). Furthermore, LISFLOOD simulates lakes and reservoirs as points on the channel network. It is not recommended to extract simulated time series at the same pixel where the reservoir or lake is located but to either extract the time series on the upstream or downstream pixels of lakes and reservoirs (depending on the location of the gauging stations for observations). More info can be found on the model documentation of LISFLOOD (https://ec-jrc.github.io/lisflood/ ). The location of lakes and reservoirs on the EFAS grid can be found also on the EFAS map viewer (https://www.efas.eu/efas_frontend/#/home ) .

**Reply:** *Thank you for highlighting the need to (1) clarify the method we employed to match the observational GRDC sites with the corresponding EFAS grid cells; and (2) assess the correspondence between the catchment area of the GRDC sites and the upstream area of the corresponding EFAS grid cells.*

*As we were working with a large data set to start with (>1500 GRDC catchments), we were not able to manually identify EFAS pixels for each of the GRDC stations in the initial data pool. Because manual matching seemed infeasible for such a large data set, instead we identified one grid cell per GRDC catchment using latitude-longitude (or coordinate) matching. As indicated by the reviewer, not all of these pixels may necessarily correspond to the 'correct' pixel with the same upstream area as the GRDC catchment. To avoid including catchments with a mismatch between upstream pixel area and GRDC catchment area, we have now pre-filtered the catchments and only included those catchments which showed a relative difference in catchment area between upstream pixel area and GRDC catchment area of < 20% in the initial catchment pool. Using this dataset (that only included catchments with a good area match), we then applied the model evaluation process, which aimed to filter out any additional catchments where simulation performance with respect to high flows was not considered sufficient for our flood frequency analysis. This two-step process (area correspondence verification and performance evaluation) allowed us to select a data set of 234 clearly-located catchments with good model performance in terms of high flows. We updated the respective passages in the manuscript to clarify the two-step procedure.*

Finally, we have found several data quality issues with the observed discharge data in GRDC in the past. We recommend strongly to have at least a visual check of the observed data that is selected for the analysis.
**Reply:** *Thank you for pointing out the need to quality check observed discharge data downloaded from the GRDC. We visualized both the observed and simulated time series of the 234 catchments we used for the frequency analysis and did not detect any obvious inconsistencies in the observed data.*

**Minor comments:**

Chapter 2.1, page 3: EFAS 4.0 as well as EFAS3.0 have been calibrated using Kling Gupta efficiency and not NSE. For more details on the EFAS versioning and what changes are included in each EFAS version please see here for a detailed description: https://confluence.ecmwf.int/display/COPSRV/EFAS+versioning+system The reference to Smith et al. refers to previous and outdated EFAS model versions that is not available on the C3S CDS.
**Reply:** *Thank you for highlighting that the calibration procedure described in Smith et al. does not refer to the most recent calibration setup. We updated the description of the calibration procedure following information provided in the EFAS Wiki (https://confluence.ecmwf.int/display/COPSRV/Modelling+upgrade+for+EFAS+v4.0).*

Chapter 2.2, page 5, line 109: "…model stability…": Please describe what is meant here with model stability? In terms of what?
**Reply:** *We have rephrased the sentence to explain what we mean by model stability: 'Next, we assess the suitability of the perturbed ensemble streamflow simulations for ensemble pooling by evaluating whether individual simulation runs can be considered independent and whether the model is stable, i.e. simulated distributions are stable across lead times (Kelder et al. 2020).'*

Chapter 2.2, page 5: It is stated that Spearmans rank correlation can only be computed for AM and not directly for POT (lines 114-116). However, in the following sentence you write that you calculated Spearmans correlation for POT. Please clarify!

**Reply:** *Thank you for pointing out the need for clarification. We rephrased the section to: 'Note that such correlation can directly only be computed for AM and not for peak-over-threshold (POT) series because POT series may differ across ensemble members in the number of events chosen for analysis and not just in timing and magnitude. It can therefore be assumed that the POT events used in our subsequent analyses are more independent than AM events. To illustrate this, we indirectly compute Spearman's correlation for pairs of POT time series by using events where at least one of the time series exceeds a threshold and by replacing non-exceedances in the other time series by 0 (not ideal because this might artificially introduce some sort of dependence).*

Chapter 2.3, page 8, line 180: … the the…. Please correct.
**Reply:** *We corrected this typo.*

Chapter 3.1, page 8: This is a repetition of Chapter 2.2. Please remove Chapter 3.1!
**Reply:** *Good point, we removed Section 3.1.*

Chapter 3.3, page 13, lines 269-271: Please describe the evidence for claiming that relative differences between simulated and observed best estimates and uncertainty bounds are independent of model performance.
**Reply:** *We specified that by 'independent', we mean 'uncorrelated'. That is, we computed the correlation between relative differences of simulated and observed best estimates and different model performance metrics (e.g. KGE).*

Chapter 3.3, Fig. 11: I disagree with your statement that flood quantiles are positively related to mean precipitation. According to Fig. 11 there is only a very weak positive relation.
**Reply:** *It is correct that the regression coefficient for precipitation is weaker than the one for slope or latitude but the explanatory variable is still significant and positively related to flood magnitude. We therefore think that this statement is correct.*

Chapter 3.3, Fig. 11: Please explain why there is such a strong positive relation to Latitude according to Fig. 11? This is not mentioned at all in the text.
**Reply:** *We agree that it is difficult to argue why latitude should physically be an important predictor of flood magnitudes. We therefore excluded longitude and latitude as potential predictors and redid figure 11. We now see a relatively strong relationship between temperature and flood magnitude. That is, higher flood magnitudes for catchments with colder climates (e.g. those in the Alps).*