Response to RC1 for hess-2021-539: Matthews, G., et al. Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System

We thank the reviewer for their thoughtful comments and suggestions which we believe will greatly improve the manuscript. The reviewer's comments have been summarised and numbered for clarity. The authors responses are in blue. Comments regarding typos or grammatical mistakes are discussed last.

General comments:

1. "The discussion of the results is, however, a bit too long and could be summarized more concisely."
We thank the reviewer for this suggestion. If invited to revise the paper, we will shorten this section. Please see the responses to comments 5, 6, and 24 which explain how this will be done.

2. [Section 2] "There are many EFAS papers available now and all describe the EFAS system. This could be shortened and only the differences to the operational settings of EFAS should be explained, in particular how the reforecasts are used."
We thank the reviewer for this suggestion. We will shorten the material currently in section 2 by focusing on the details that are different between the operational forecasts and the reforecasts and removing the additional details. For example, we will remove the details of the NWP systems and direct the reader to the relevant sources but will discuss that the operational forecasts are multi-model ensembles whereas the reforecasts are from a single model. To avoid repetition, we will move some material from the current section 4.2.1 that discusses the reforecasts to section 2.

3. "This small number of members causes difficulties in computing the CRPS for making a fair comparison with a CRPS derived from the PDFs of the post-processed forecasts (see for example Zamo and Naveau, 2018). However, this problem is not mentioned and the presented results of the CRPSS should be treated with care."
We thank the reviewer for this comment and for highlighting the relevant literature. If invited to revise the paper, we will add the following to section 4.3.4: "It should be noted that the error in the calculation of the CRPS for the raw ensemble forecasts is likely to be large compared to that of the post-processed forecasts because of the limited number of ensemble members (Zamo and Naveau, 2018). However, as this evaluation is of the post-processing method no corrections to account for the ensemble size are made (e.g. Ferro et al., 2008) as the impact of the post-processing would be difficult to differentiate from that of the CRPS correction."

4. "A most recent paper from Skøien, et al. (2021) has a similar topic about evaluating the post-processing methods for EFAS (EMOS and the application of transformations like NQT). Therefore, the differences and novelty of this study should be stressed clearly and discussed in more detail."
We thank the reviewer for bringing this relevant article to our attention. The Skøien et al. (2021) study differs from our study in several ways: (a) Skøien et al. research the implementation of a new post-processing method whereas the post-processing method that we use is the operational post-processing method used in the EFAS system. Therefore, our results are of interest to end-users of the EFAS forecast products as well as the hydrological research community. (b) The focus of Skøien et al. is to compare different methods with a focus on forecast features whereas we compare a single method across many catchments with a focus on catchment characteristics. (c) The combination of the MCP and the EMOS methods allows us to account for the errors in the hydrological model as well as in the meteorological forcings whereas Skøien et al. did not investigate the hydrological model uncertainty. (d) Skøien et al. use lead-time dependent EMOS parameters whereas we use lead-time invariant parameters. We will discuss points a and b in the section 1 and points c and d in section 5.2.1.

5. "Although the analysis of the different aspects like catchment size, elevation, regulation, length of period, are very interesting, it maybe could be shortened focusing on floods, which is the main topic of EFAS."
We thank the reviewer for this suggestion. As more end users are using EFAS forecasts, particularly the post-processed forecast products for which no flood alerts are issued, for information about non-extreme streamflow situations as well as for floods, we believe that the evaluation at all flow levels is beneficial. However, we will

shorten sections 5.2.1-5.2.3 by restructuring each section to remove repetition. For example, lines 1028-1032 and lines 1043-1045 make similar remarks so we will combine these comments.

6. “Also, the detailed analysis of the Nash-Sutcliffe (KGE) is maybe too long, since the results of the post-processing methods are probabilistic and the KGE reduces the information content to the mean (median) of the Ensembles (or PDF).”
We will shorten section 5.1.1 by removing some of the specific quantitative detail. For example, on line 622 we will remove "where approximately 30\% of stations have an improved correlation of over 0.1”. We will also remove repetitive descriptions of Fig. 5 (e.g. line 620). Additionally, we will reduce the discussion on the drift in β-values with lead-time by focusing only on the impact of non-stationarity.

Specific comments:
7. [Page 5, section 2] “Highlighting the differences between the operational setting of EFAS and the setting used in this study should be sufficient. More details can be found in many other papers.”
Please see the response to comment 2.

8. “However, the calibration period of the LISFLOOD model is missing. Is there an overlap between the period for calibrating the parameters of the hydrological model and the historical period p for the off-line calibration?”
The calibration of LISFLOOD was performed over the period 1990 – 2017 using 6-hourly data where available (Mazzetti et al., 2021). Therefore, there is an overlap with the historic period, p. We will add this information to Section 2.

9. [Figure 1] “the index of the parameters μ and Σ is ψ, whereas in the caption you use the index Φ.”
We thank the reviewer for highlighting this inconsistency. If invited to revise the paper, we will change the index of parameters μ and Σ in the caption of Figure 1 to ψ.

10. “You mention several times (e.g. line 172, page 7) that the minimum for the off-line calibration is 2 years. However, for the fitting of the GPD you use 1000 values (page 9). So you will need more than 2 years?”
If the length of the calibration timeseries for a station is only two years (less than 1000 data points) then the station is still calibrated but all data points will be tested as the break point. We will add this clarification to Section 3.3.1.

11. “I would suggest including a list of nomenclature, so you avoid repetitive descriptions like the tilde for the physical space (line 182, 204, 220, 340) and the definition of the timestep notation introduced on page 7 (lines 176-177).”
We will remove the repetitive descriptions and direct the readers to the notation section (Section 3.1).

12. [Line 172, page 7] “you write that the location parameter a is used for defining the breakpoint, but in Figure 2 the shape parameter c is used as breakpoint.”
We thank the reviewer for highlighting this inconsistency. We will change the variable name for the breakpoint in Figure 2 to a.

13. [Line 251, page 9] “you write about consistence between the 2 distributions. What does it mean? How do you check this?”
To create a smooth discharge distribution there must be no step change in the distribution at the breakpoint between the kernel density estimation and the Generalised Pareto Distribution (GPD). We will replace line 251 with “The scale parameter, b, is determined analytically by the constraint that the density distribution must be equal at the breakpoint for both the GPD and the KDE distributions.”

14. [Line 257, page 10] “the concentrated likelihood method is mentioned without any further explanation what this method does. Maybe some more details would be helpful.”
The concentrated likelihood method requires one parameter, say a, to be expressed as a function of the other parameters, say b and c. The likelihood of a can then be calculated given the values of b and c. We will explain the concentrated likelihood method and how it is applied within the method.

15. [Lines 261-262] “Also, it is not clear for me how the GPD is weighted”

Given that the total probability density must equal 1, the weighting of the kernel density part of the distribution is the value of the cumulative distribution function at the breakpoint, denoted F(a) and the weighting of the GPD part of the distribution is then 1-F(a). We will change lines 261-262 to "The likelihood function for the full distribution is the product of the likelihood function of the kernel density and the likelihood function of the GPD weighted by their contribution to the total density, F(a) and 1-F(a), respectively (MacDonald et al., 2011)."

16. [Lines 270-280, page 11] "Whereas the description of the linear approximation is maybe not necessary."
We thank the reviewer for this suggestion. We will remove the detailed description of the linear approximation.

17. [Page 13] "it is not clear for me why you have observations (line 336) and water balance (line 337) for the period until t, but the forecasts (line 339) until t-1?"
We thank the reviewer for highlighting this ambiguity. The forecast that is produced at time *t* is the forecast that we are post-processing. Therefore, although observations and water balance values are available (in this idealised situation) for time *t* and can be used in the post-processing method, the forecast at time *t* cannot. To clarify this issue we will change line 335 to "As well as the current forecast produced at time t, $x_t(t+1 : t+T)$, the online correction requires the following input data from the recent period:".

18. [Line 411, page 16] "you write that a set of forecasts are used to estimate the two spread correction parameters. How did you choose the size of these sets of forecasts?"
The size of the set of forecasts is the number of forecasts available within the recent period. The length of the recent period, q=40, was determined using tuning experiments that were performed prior to the work of this study (Paul Smith, personal communication, September 25, 2020). We will explain this in the introduction to Section 3.4.

19. [Figure 3] "In Figure 3 you write CRPS besides the legend bar, but it should be CRPSS?"
Figure 3 shows the CRPS of raw reforecasts with the water balance used as the "true" value. We use this metric as a measure of the meteorological error which is discussed within the results in Section 5.1.1. We will sign-post the reader to this results section to motivate the use of the CRPS.

20. [Line 483, page 19] "you write that only 11 reforecasts are available (I suppose this 11 comes from 40 days/7day x 2). The number 11 could be misleading, since it happens to coincide with the number of members mentioned in the next sentence"
We thank the reviewer for highlighting this ambiguity. We will replace line 483 with: "Whereas operationally daily forecasts for each day of the recent period are available, here only two reforecasts are available for each week of the recent period reducing the number of forecasts used to estimate the EMOS parameters from 40 to 11."

21. "Why do you fix it [q] to 40 days? Since there is this a discrepancy between the operational setting and this analysis anyhow, you could set q to a longer period to include more reforecasts (e.g. q=70 ~ 20 reforecasts)."
We thank the reviewer for this comment. We chose not to extend the recent period to account for the reduced number of reforecasts to avoid seasonal variations in the EMOS parameters. will explain this in section 4.2.1.

22. [Line 486] "I don't understand why the mean discharge value is predicted for the previous 6 hours?"
In the EFAS 4 system the hydrological model, LISFLOOD calculates the discharge as the average over the previous 6 hours for both the water balance simulations and the forecasts. However, the post-processing is currently only performed at a daily timestep. Therefore, the 6-hourly modelled discharge values must be aggregated to a daily timestep. We will add this explanation to Section 4.2.1.

23. [Figure 11a] "The difference between the raw and the post-processed forecasts in Fig. 11a (mentioned on page 35, line 862) is very difficult to see and almost not visible."
We thank the reviewer for highlighting this issue. Line 862 discusses two peaks which are forecast by the raw forecasts but not the post-processed forecasts. If invited to revise the paper, we will highlight these two peaks using boxes (see Figure 1 below).

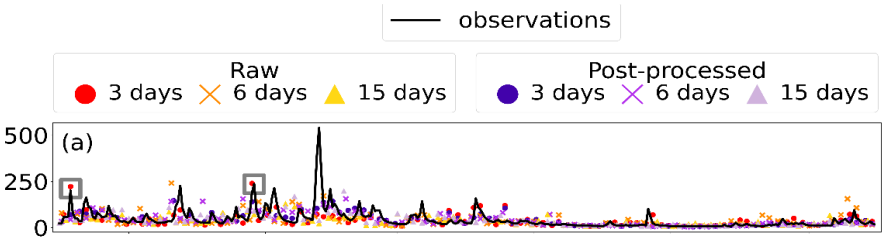

Figure 1: Modified version of Figure 11a. The grey boxes highlight the two peaks discussed on line 862.

24. [Page 41] "the paragraph from line 1004 – 1010 can be removed"
If invited to revise the paper, we will remove this paragraph.

25. "I have some doubts about your suggestion that very short periods are sufficient (line 1045): the chance that such a short period will show the variability of the discharge needed for applying the NQT is rather small and the fitting of the GPD almost impossible. Consequently, the back-transformation of the variables from the Normal space will always produce poor and very unreliable results for floods."
We thank the reviewer for this comment. We will change the wording at line 1045 to "Although a full sensitivity analysis is beyond the scope of this study, these results suggest that very short time-series can be used, if necessary, to correct for consistent biases, although longer time-series are preferable. However, care should be taken when forecasting high flows since a short timeseries will not allow for a robust calculation of the upper tail of the discharge distribution (see Sect. 3.3.1) which will likely cause errors in the forecast probability distribution (Bogner et al., 2012)."

26. "The citation of Coccia (line 1135) is incomplete."
We thank the reviewer for this suggestion. We will correct this reference.

27. "Also, the term "Multi-Temporal" in combination with the MCP (MT-MCP) is mentioned only in line 153-154 and in the conclusions (line 1054), but is not explained."
We thank the reviewer for highlighting this missing explanation. We will add the following to the introduction (line 64): "The method discussed in this study is partially motivated by the Multi-Temporal Model Conditional Processor (MT-MCP, Coccia, 2011) which extends the original MCP method for application to multiple lead-times simultaneously."

28. Grammatical errors and typos
    a. "Line 349: …using a MCP method …"
    b. "Page 14, line 372: in the recent period …"
    c. "Line 880: …uncertainties show a small increase"
    d. "Line 1021 ..greater than.."

We thank the reviewer for highlighting these errors. We will correct these errors.

References:

Zamo, M., Naveau, P. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. Math Geosci **50,** 209–234 (2018). https://doi.org/10.1007/s11004-017-9709-7)

Ferro, C.A.T., Richardson, D.S. and Weigel, A.P. (2008), On the effect of ensemble size on the discrete and continuous ranked probability scores. Met. Apps, 15: 19-24. https://doi.org/10.1002/met.45

Skøien, J. O., Bogner, K., Salamon, P., & Wetterhall, F. (2021). On the Implementation of Postprocessing of Runoff Forecast Ensembles, Journal of Hydrometeorology, 22(10), 2731-2749

Mazzetti, C., Decremer, D., Prudhomme, C. Challenges of the European Flood Awareness System (EFAS) hydrological calibration. Poster presented at: Joint Virtual Workshop on "Connecting global to local hydrological modelling and forecasting: scientific advances and challenges"; June 29, 2021; Online [Available: https://events.ecmwf.int/event/222/overview].

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G.: A flexible extreme value mixture model, Computational Statistics & Data Analysis, 55, 2137–2157, 2011.

Bogner, K., Pappenberger, F., and Cloke, H.: The normal quantile transformation and its application in a flood forecasting system, Hydrology and Earth System Sciences, 16, 1085–1094, 2012.

Coccia, G.: Analysis and developments of uncertainty processors for real time flood forecasting, Ph.D. thesis, alma, http://amsdottorato.unibo.it/3423/, 2011.