Response to RC2 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 helpful comments and believe that they will make the evaluation more rigorous and useful. The reviewer's comments have been numbered for clarity. The authors responses are in blue.

Full review:

1. "All aspects presented are of interest, however I do wonder whether the paper could be separated into two more focussed manuscripts, perhaps one focussing on the novel aspects of the post-processing method and validating its assumptions, and a second on evaluation the benefits and investigating factors that influence its performance."

We thank the reviewer for this suggestion. Although we had considered separating the paper into two papers, we prefer to keep the manuscript as one paper, as we feel it is important that the methods are discussed alongside their practical performance. However, we plan to shorten the paper (see response to RC1, particularly comments 5 and 6).

More specific comments:

2. "The sample covariance matrix is used to characterise the joint distribution of the historic observations and water balance simulations, equation 7. There are potential issues that may be encountered using this approach and it would be good understand whether special treatments have been needed to overcome these."

   i. "The covariance matrix is computed over a set of historic observations and is likely to have inflated, or spurious, correlations over long lags if the seasonal cycle of streamflow is not considered."

Spurious correlations are not treated in the current method and the joint distribution is naively assumed to be consistent throughout the year. The main reason for this is that many stations do not have a sufficiently long timeseries to consider seasonal distributions. If invited to revise the paper, we will add this information to Section 3.3.2.

   ii. "For large sample covariance matrices such as those estimated in this study, missing observations can lead covariance matrices that are not positive definite."

The covariance matrices are adjusted using a minimum eigenvalue threshold approach to guarantee they are positive definite. We will include this process in Section 3.3.2.

3. "The KGE analysis is performed using the median as a point estimate of the forecast ensemble. The results obtained for the post-processed forecasts, particularly the bias ratios and variability ratios of less than one at long lead times, are not unexpected as the variance of the forecast median will be considerable more damped that the mean. The forecast mean is likely to be a better choice as the point estimate of the forecast ensemble. Some theoretical justification of the use of the ensemble mean with measures of squared error can be found in Gneiting (2011)."

We thank the reviewer for this suggestion. We have now performed the calculations of the modified KGE for the ensemble mean. Figure 1a (below) shows the modified KGESS for the ensemble median (blue) and the ensemble mean (green) and shows that the distribution of the KGESS is similar for both point estimates. Additionally, Figures 1b, 1c, and 1d show the components of the modified KGE for the ensemble median (orange), post-processed forecast (purple), and the ensemble mean (cyan). As we can see in Fig 1b and Fig. 1d, the distributions of the correlations and variability ratios are similar for the ensemble mean and the ensemble median. In Fig. 1c we see that the median of the distribution

of bias ratios (shown by the central black line) for the ensemble mean (cyan) is higher at all lead-times than those of the ensemble median and the post-processed forecast.

In the original manuscript, the ensemble median was chosen because operationally the ensemble forecasts are often represented by boxplots where the median at each timestep is shown. We believe that the comparison of the ensemble median with the median of the post-processed forecast is a useful evaluation for end-users who may be choosing between the two products. Therefore, we will add text motivating the choice of the ensemble median in Section 4.3.1. However, we would also briefly discuss the ensemble mean and include the lower panels of Figure 1 (1b, 1c, and 1d) in the supplementary material.

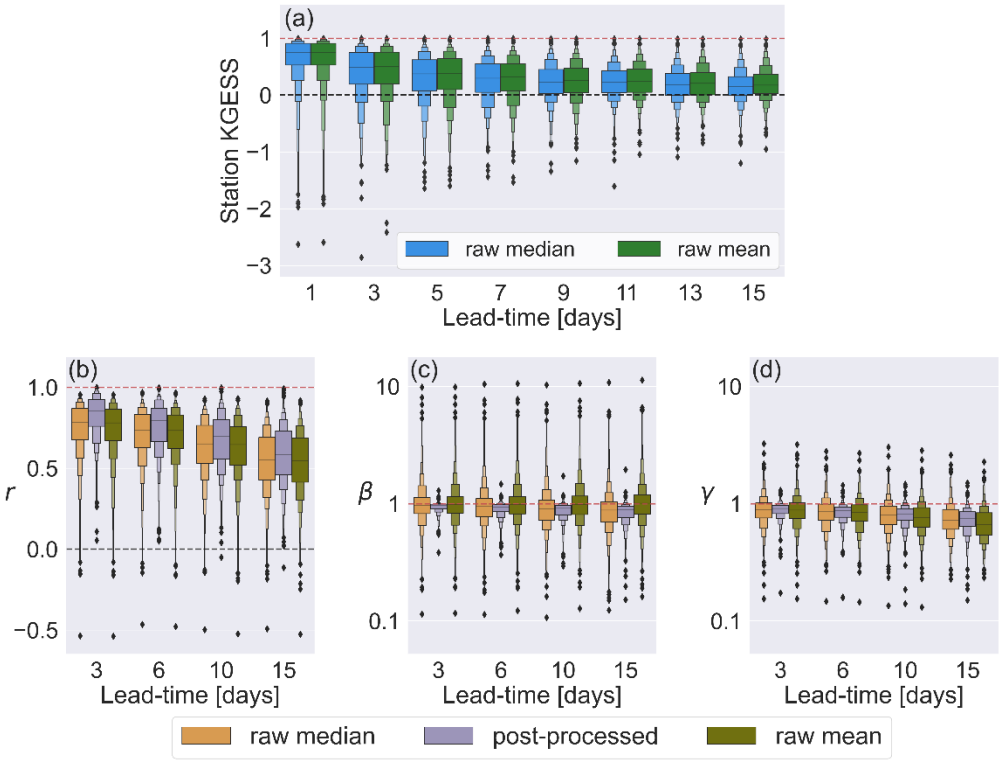

Figure 1: Kling-Gupta Efficiency analysis for the raw ensemble median, post-processed median, and the raw ensemble mean. (a) KGE skill score with the post-processed forecast as the benchmark, (b) correlation coefficient, (c) bias ratio, (d) variability ratio.

4. "In this paper, the analysis of peak timing is conditioned on observations exceeding a threshold (90th percentile discharge threshold) within the forecast period and is likely to result in a biased evaluation of forecasts. A more rigorous approach would be to select the events based on forecasts exceeding the threshold."

We thank the reviewer for highlighting this limitation. We have run the analysis again using a forecast exceedance threshold and note that the bimodal distribution that was shown for lead-times 10-15 days is no longer present (see Figure 2 below). We believe the bimodal nature was due to forecasts that failed to predict an event being more harshly penalised than forecasts that predicted an event that did not occur. We will change the peak-time error analysis to use a forecast exceedance threshold.

5. “I also believe that rather than evaluating the timing of the peak in the forecast median, which doesn't correspond to the peak in any individual hydrograph, a more representative point estimate of the forecast timing error would be to compare the median (or mean) time to peak across all ensemble members to the timing of the observed peak. “

We thank the reviewer for this suggestion. We have performed the analysis with both the peak of the ensemble median and the median of the peaks of each of the ensemble members and found that the distributions are very similar. We have rerun the analysis using the criteria of forecast exceedance (see reply to comment 4). Figure 2 (below) shows the peak-time error for the peak of the ensemble median of the raw forecast (orange), the post-processed median (purple) and the median of the peaks for all ensemble members of the raw forecast (green). The two distributions calculated from the raw ensemble forecasts are similar in comparison with the post-processed forecast with the raw median performing slightly better at longer lead-times. Therefore, we will keep the comparison between the post-processed forecast and the ensemble median, but we will use the exceedance of the forecasts as the event criteria (see reply to comment 4).

6. “line 373 - values in the recent perion should be "values in recent period"

We thank the reviewer for highlighting this mistake. We will correct this mistake.

7. “Line 825 - CRPS calculated on deterministic forecasts is equivalent to the absolute error not the square absolute error.”

We will also correct this mistake.

8. “Figures - The size of multi-panel figures (e.g. Figure 9, 12) could be increased to better illustrate the detail”

We thank the reviewer for highlighting the figures which are unclear. We will increase the size of Figures 5, 6, 9, and 12.

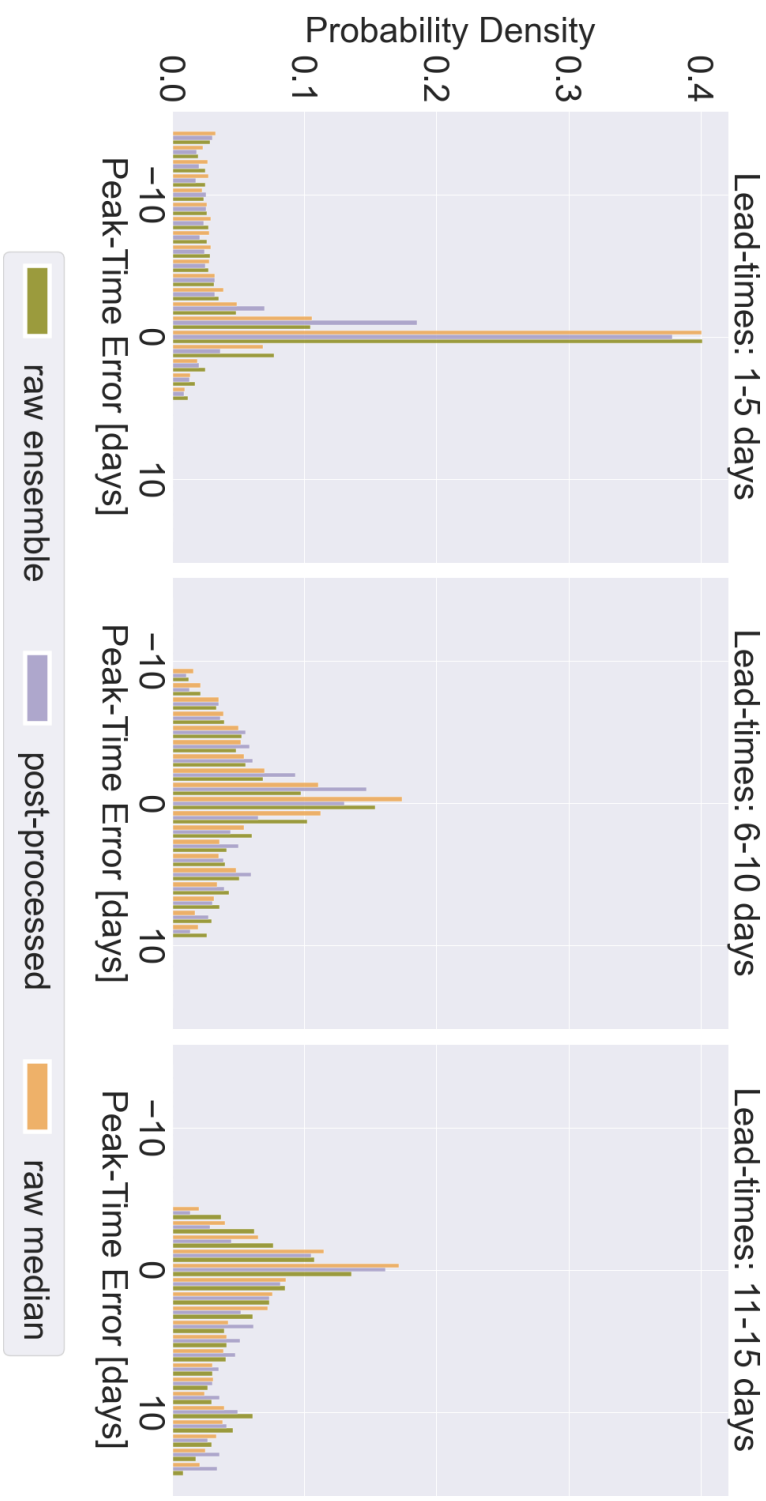

Figure 3: Peak-Time Error analysis using forecast exceedance as the event criteria for the peak of the raw ensemble median(orange), the peak of the post-processed forecast median (purple), and the median value of the peak timing of the ensemble members (green)