Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-539-RC2, 2022

Comment on hess-2021-539

Anonymous Referee #2

Referee comment on "Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System" by Gwyneth Matthews et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-539-RC2, 2022

This paper describes a novel post-processing method applied to the EFAS forecasts, assesses the improvements to forecasts realised by post-processing to a large number of catchments and investigates factors that influence the performance of the post-processor. The paper is very well structured and written, and the topic is of considerable interest to the forecasting researchers and practioners using the EFAS forecasts.

The paper is comprehensive covering the post-processing method itself, the improvements across the EFAS domain and the factors influencing the forecast performance, which necessitates a lengthy manuscript. All aspects presented are of interest, however I do wonder whether the paper could be separated into two more focussed manuscipts, 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.

More specific comments:

The sample covariance matrix is used to characterise the joint distribution of the historic observations and water balance simulations, equantion 7. There are a potential issues that may be encountered using this approach and it would be good understand whether special treatments have been needed to overcome these. Specific issues that come to mind include: (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. These inflated/spurious correlations are likely to lead to inflated variances of conditional predictions. (ii) The authors indicate that there missing (and possibly zero-valued) observations that are used in the estimation of the covariance matrices. For large sample covariance matrices such as those estimated in this study, missing observations can lead covariance matrices that are not positive definite. Have any issues been identified and any special treatment been implemnted to deal with

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).

Analysis of forecasts for extreme events such as floods requires careful design to ensure that the performance evaluation is not biased (Lerch, 2017). 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 rigourous approach would be to select the events based on forecasts exceeding the threshold. 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.

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

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

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

References:

Tilmann Gneiting (2011) Making and Evaluating Point Forecasts, Journal of the American Statistical Association, 106:494, 746-762, DOI: 10.1198/jasa.2011.r10138

Sebastian Lerch. Thordis L. Thorarinsdottir. Francesco Ravazzolo. Tilmann Gneiting. "Forecaster’s Dilemma: Extreme Events and Forecast Evaluation." Statist. Sci. 32 (1) 106 - 127, February 2017. https://doi.org/10.1214/16-STS588