Response to reviewer #2

We thank the reviewer for their careful reading of and thorough comments on our manuscript. In the following, we repeat the reviewers' comments for clarity and added our replies to them in italic font. Additions and changes to the manuscript are indicated by an italic and bold font.

I would like to thank the editor for the opportunity to review this manuscript.

This study investigates the performance of three continental- / global-scale streamflow forecasting systems (HTESSEL, EFAS and E-Hype) in predicting inflows to Lake Ijssel – a major surface water reservoir in the Netherlands. All three forecasting systems are driven with ECMWF SEAS5 seasonal climate forecasts but differ in their underlying hydrological modelling approach. The authors applied bias correction to streamflow forecasts using Quantile Mapping and subsequently assessed the skill of raw and post-processed forecasts for predicting inflows. A particular focus was placed on dry conditions, i.e. the predictability of low-flow events.

In my view, the overall focus of the study to compare three leading continental- or global-scale streamflow forecasting systems for predicting streamflow at the local-scale is interesting and I believe the study has applications and implications that may be relevant to a wider audience, e.g. how to best translate outputs from continental- or global-scale streamflow forecasts into local-scale applications. However, my main concern relates to the novelty of the study. In its current form, the manuscript is written similar to a Technical Report that describes methods and results of forecast post-processing and verification applied to one single study location, without sufficiently linking it to the research context, and may be of interest to a very local audience. I am missing an attempt to generalise the findings and place them into a broader context and emphasise implications that are applicable in other regions too (using Lake Ijssel inflows as a case study). If the authors addressed this, I believe the study would be of interest to a wider audience.

The study design and methodology (including post-processing and forecast verification) are robust. While more advanced streamflow post-processing methods exist and could be investigated, the authors clearly show an improvement in skill for the study location of interest. The application of multiple verification metrics (continuous ranked probability skill score, mean error, Brier skill score and reliability diagrams) to 23 years of hindcast data, in a cross-validation approach, is appropriate and thorough.

Major comments:

Novelty and relevance to a wider audience: As outlined above, the novelty of the study and relevance to a wider audience is not very clear. I would ask the authors to place the work into a broader research context and interpret the results more broadly, highlighting implications for researchers and practitioners in other regions. While the introduction places the research into a wider context to some extent, there is no clear link between the introduction and the rest of the manuscript which describes the results in a very detailed way, focusing on one individual location. It would be great if the authors could come back to the broader research gap in their discussion and conclusions, and interpret the results more broadly, e.g. what are key messages for the hydrological community who aim to apply regional- or global-scale seasonal streamflow forecasts for individual catchments? How does this study compare to similar systems implemented in other parts of the world? What are advantages and disadvantages of the described approach?

Results and Discussion: The results section is very comprehensive and describes the results of the forecasting approach in a very detailed way, focusing on a range of forecast performance metrics. However, I am missing an interpretation that goes beyond simply describing the results. The discussion section itself is very short and immediately starts with limitations of the study and further research, without an actual discussion and interpretation of the study findings. I would ask the authors to add more discussion of their results – linking them back to their research aims or questions that are outlined in the introduction, placing them into the context of the existing literature and highlighting implications or applications of the findings (followed by limitations and further research, as is already included). Potential points of discussion: What can we learn from these results that is relevant in other regions? Can some of the differences in results be traced back to differences in the underlying specifications of the hydrological models (e.g. consideration of routing) and what does it mean for other locations? Could a multi-model ensemble approach be useful?

Readability: I had difficulties following some sections. The manuscript would benefit from revision with the aim of re-wording sections and sentences to be clearer and more concise. Additionally, there are some typos throughout the text (I have included a few examples under “minor comments” but it applies to the manuscript overall).

We agree with the reviewer that the discussion could be more extensive and thank also this reviewer for the suggestions. We substantially extended the discussion and in effect added a number of paragraphs to the discussion section, discussing the results in a broader context.

We agree with the reviewer that the discussion could be more extensive and thank him/her for the suggestions. We substantially extended the discussion and in effect added a paragraph to the discussion section, discussing the results in a broader context:

Our results largely confirmed earlier results by (Arnal et al., 2018), who found increased forecast skill in spring in (partly) snowfed rivers and lowest forecast skill in summer. In winter, teleconnections like NAO positively affect forecast skill in European rivers (Scaife et al., 2019; Bierkens and van den Hurk, 2007). Arnal et al. (2018) indicated few regions in Europe with increased forecast skill in summer, but the Rhine appeared not to be one of them. To ensure statistical robustness, these analyses resulted in average forecast skill over a large number of years. Although we confirmed that average summer

forecast skill is low and varies between years, we also found that skill increases with the extremity of the event: summers with extremely low discharges were skillfully forecast longer ahead, up to four months. Ionita and Nagavciuc (2020) found similar results for the summer of 2018 based on statistical forecasting systems. They found especially sea surface temperature in parts of the northern Atlantic ocean to be a good predictor of Rhine river discharge for long lead-times. Meißner et al. (2017) found such statistical forecasting methods to outperform more physically based methods such as SEAS5. The forecast systems used in this study all depend on meteorological forecasts based on the ECMWF model. Recently, multi-model forecast systems have been developed using different general circulation models (GCMs) for atmospheric forecasts (Samaniego et al., 2019; Wanders et al., 2019; Muhammad et al., 2018). This highlights the importance of multi-model forecasting systems incorporating both statistical and physically based methods.

Our finding of increasing predictability with event extremity suggests that individual forecasts contain useful information, which could be discarded as noise by statistical analyses. As was also noted by Viel et al. (2016) and Meißner et al. (2017), small forecast skill does not mean forecasts are not useful for decision-makers. Our results indicate that when a large fraction of the forecast ensemble is in the lower or upper tercile, the probability of the forecast being correct is high. This, therefore, is very important information for decision makers in the reservoir management in anticipation of extreme low-flow conditions.

Given the identical meteorological forcing, the differences found in our study are presumably caused by the schematization and/or parameterization of the hydrological models. By considering and comparing the relevant processes separately (e.g. groundwater flow, glacier melt) the reasons for these differences could be further explored. The representation of glacial melt, for example, is more sophisticated in E-HYPE compared to Lisflood, which might contribute to slightly higher skill in summer for E-Hype, whereas the opposite may be the case for snow melt, causing slightly higher skills for EFAS in early spring. The absence of hydraulic processes in H-TESSEL is likely to cause overall lower forecast skills. Fully disentangling these differences requires model output per component for all forecast systems, which was not feasible in the current study.

We focused on one specific location in the Rhine basin. However, discharge at Lobith nearly integrates the entire Rhine basin, which covers, with 160,000 km2, a substantial part of Western-Europe. In more spatially extensive studies, (e.g. Arnal et al., 2018), results for the Rhine appeared comparable to many other areas in Europe, suggesting that our results are applicable to other areas in Europe. In general, streamflow is more predictable in river systems with long memory due to snow-processes, groundwater contribution and dampening from lakes and reservoirs and groundwater contribution, all of which apply to the Rhine, and less in arid climates with fast hydrological response to precipitation (Pechlivanidis et al., 2020). Our finding of the higher predictability of low flow extremes would, in our view, also translate to other catchments with similar characteristics. This remains to be confirmed in a future study.

Regarding the readability of the manuscript, we re-read the manuscript and clarified the wording where we deemed appropriate to do so. This affects a number of locations, which we will not all repeat in this letter.

Other comments / suggestions:

Some of the results and figures are very detailed and could be presented in a more concise or synthesised way. A few suggestions and observations in relation to the figures:

Figures 4-6: Could you present the results for each forecasting system (raw and bias corrected outputs) and each lead time, aggregated over all months (April to September), to be able to compare the systems more generally? One possibility could be to add an additional sub plot that presents aggregated results across all months.

We added a panel to these figures showing the skill scores aggregated over all shown months. Concerning the more general remark about a more concise presentation of the figures: we believe the month-by-month analysis of forecast skill is essential for the storyline. We, therefore, chose to add an extra panel. We also added result for March, to keep an even number of panels.

Figure 10: Similar as above, it would be great to see the overall RPSS scores aggregated over all years and months, to compare the three systems directly with each other. Would it be possible to add another sub plot that presents the results aggregated across all years and months?

This figure is different compared to the above as it is intended to show the variability between years. Aggregated over all available years, it would show similar results as were shown in Figures 4-6 (albeit with slightly higher skill as we only consider 'extreme' years). Moreover, we do not have all the relevant years available for E-HYPE, so we cannot directly compare the aggregated results.

Figure 11: I found it a little unusual to see the years on the x-axis, but in order of dryness / wetness. Would it be possible to use annual streamflow as x-axis (i.e. presenting forecast skill as a function of average streamflow)?

We agree that discharge is a more logical variable to display on the x-axis. However, we deem it important for the storyline to also show the year numbers. We, therefore, now show the average summer discharge on the x-axis and added a top axis displaying the year numbers.

Figure 12: It is not clear to me what the right column of Figure 12 shows. It would be good to provide more explanation in the text and/or caption – how could this be interpreted and used?

We agree that the information in this panel is somewhat unconventional and arose during discussions with local water managers. We rewrote the corresponding text in Sections 3.2.2 and 4.2, as well as the figure caption. We now refer to it as the absolute difference between higher-than-normal and lower-than normal discharge and extended the explanation. What the metric is intended to show is that the reliability of the forecast increases as this absolute difference increases. It is, therefore, related to a reliability plot but appeared to appeal to water managers as it can provide them directly with a probability that a forecast will turn out correct, as a function of the presented absolute difference.

Minor comments:

P2 L47: Please change "takes" to "take" and remove "also".

Done.

P3 L61: Space missing after “as follows.”

Corrected.

P3 L79: I suggest to change “data forecasting systems” to “seasonal forecasting systems”.

We agree with the suggestion; corrected as suggested.

P6 L135-136: Please change “is investigated 3.2.2” to “is investigated in Section 3.2.2”.

Done.

P8 L175: “using” instead of “usiing”

Done.