Review for HESS-2021-604: Comparing seasonal streamflow forecast systems for management of a fresh water reservoir in the Netherlands

In general, the study adresses a very recend and highly relevant topic as it discusses the skill of seasonal forecasts for predicting runoff. Besides the "raw" runoff forecasts from HTESSEL, E-HYPE and EFAS, the authors also apply a simple but effective quantile mapping bias correction in order to match the distribitions between observed and forecasted runoff. The bias-corrected forecasts are then compared against observed runoff at a single gauge in the river Rhine at Lobith, that serves as a proxy for inflow to the river and lake IJssel. The authors conclude that after bias correction, streamflow forecasts have skill up to four months ahead in spring, when streamflow is dominated by snow melt. During summer, this skillful period decreases to about 1-2 months. While the study focuses on a single basin, the results are still important for the hydrological community.

Major comments

- But this limited scope is my main poin of criticism. As far as I can tell, the authors did not run any of the models, but rather used publicly available data. Furthermore, there are multiple data sources for runoff observations across Europe. Thus, while the study is certainly of high relevance for the water management of the IJssel reservoir, it could serve as a state-of-the-art skill assessment for a much larger domain and, hence, would be an update for, e.g., Arnal et al. (2018). None of the applied methods are tailored to the study domain or require any local adjustments or knowledge. Thus, the whole framework could be easily applied to multiple sites and, hence, present the skill of seasonal streamflow forecasts from current hydrological forecasting systems for major European river basins. This would be a very important and much needed contribution to the whole European hydrometeorologcal community. Therefore, I would suggest to include at lease some more rivers and gauges to a) increase the importance and scope of the study and b) see if the author's findings can be transferred to other river basins.
- Furthermore, other climatic variables like precipitation or temperature are equally important for the reservoir management. If the authors decide to stick to the IJssel reservoir, I would strongly suggest to include some other variables as well to make the whole study a bit more comprehensive.
- While this is a quite frequent comment, I also think that the overall language, wording and presentation of the paper could be substantially improved.

Minor comments

- Line 61: Add whitespace before "In section 2..."
- Lines 88 - 89: Maybe re-phrase to: Total runoff from the HTESSEL land surface scheme is aggregated at native resolution over the Rhine catchment?
- Line 89: Could you give a reference for this statement? Could the lack of a routing scheme explain the bad results of the raw HTESSEL forecasts?
- Line 99: similar to, for example, the Variable...

- Line 104: The HYPE-model, whose European application is referred to as E-HYPE, is a semi-distributed...
- Line 106: Whitespace after 215km^2.
- Line 125: This is the first time that you discuss the runoff observations that are used in the study. Please add some more details like the source of runoff observations to, e.g., section 2.2.
- Lines 149 - 150: How did you compute the CRPSS for precipitation forecasts? Which "reference" did you use?
- Lines 156 - 158: Remove "three in this case" as this is obvious from the previous sentences.
- Line 158: ...is benchmarked using AN observed streamflow climatology...
- Lines 161 - 163: This sounds extremely complicated and, to some extend, different to the general definition of the term "forecast resolution". But this might be unintentional and could be improved by simply citing standard references for the evaluation of forecasts.
- Line 171: lead tme
- Line 175: usiing
- Lines 185 - 192 and Figure 7: I do not know how "substantial" this Diebold-Mariano-Tests really are to your overall asssesment... Especially as you are comparing ensemble medians and, hence, do not consider the full ensemble. I assume that a simple comparsion of forecast-tailored skill-scores (e.g., a generalization of Figure 10) should give more comprehensive information.
- Line 201: ...QM-factors are derived DURING 2000, 2001, and 2004 - 2015
- Line 258: Why are you particularly referring to machine learning approaches here?
- Section 3.1: The method is ussually known as "Quantile mapping". Please correct this throughout the manuscript. Furthermore, how did you treat the tails of the distribution? Getting these parts of the distributions right is quite important especially as you are also focusing on low-flow events in your evaluation.
- Section 3.2: Maybe replace "Forecast skill metrics" with "Evaluation of forecast skill"?
- Section 3.2.1: Why did you only use the BS for low-flow forecasts? The skill of high-flow forecasts is quite important as well.
- Section 5: You do not really "discuss" your findings here, but rather summarize your study and give an outlook. As there are already several publications on this topic (e.g., Arnal et al., 2018; Samaniego et al., 2019; Ionita & Nagavciuc, 2020), it would be interesting to put your findings in the context of these other studies.
- Figure 1: What are the black dots? Increase fontsize of legend, R^2 and y --> hard to read
- Figure 2: Please increase font-size for the legend. Furthermore, I would remove "Forecast bias corrections using...." and rather just write "Upper row: CDFs from EFAS, E-HYPE, ... derived from

forecasts issued on April 1st for lead months 1 (April), 3 (June), 5 (August) and 7 (October); bottom row: mapping factors between observed and forecasted CDFs".

- Figure 4: Do you evaluate the skill of Apr. to Sept. forecasts from different issue dates (months) or rather the skill of forecasts that have been issued from Apr. to Sept.?
- Figure 8: You write that you're using raw HTESSEL-forecasts but in the text (Lines 198 - 201), you write that "we consider the bias-corrected results". Or did I understand something wrong here?
- Figure 9: Increase the "thickness" of the observations as they are very hard to distinguish from all other lines. Furthermore, maybe use boxplots for showing the ensemble spread from the three models. Right now, the gray lines just create a lot of "noise".
- Figure 11: Please remove the lines between the dots! You do not show a continuous time-series here!
- Figure 12: Usually, the x-axis in reliability plots shows the Forecast probability and the y-axis the observed relative frequency. Is there any reason why you have not defined the axes like this?
- Table 1: Resolut0ion should be km^2; remove bracket after dampening

References

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrol. Earth Syst. Sci., 22, 2057–2072, https://doi.org/10.5194/hess-22-2057-2018, 2018.

Ionita, M., Nagavciuc, V. Forecasting low flow conditions months in advance through teleconnection patterns, with a special focus on summer 2018. Sci Rep 10, 13258 (2020). https://doi.org/10.1038/s41598-020-70060-8

Samaniego, Luis, Stephan Thober, Niko Wanders, Ming Pan, Oldrich Rakovec, Justin Sheffield, Eric F. Wood, Christel Prudhomme, Gwyn Rees, Helen Houghton-Carr, Matthew Fry, Katie Smith, Glenn Watts, Hege Hisdal, Teodoro Estrela, Carlo Buontempo, Andreas Marx, and Rohini Kumar. " Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe", Bulletin of the American Meteorological Society 100, 12 (2019): 2451-2472, accessed Feb 22, 2022, https://doi.org/10.1175/BAMS-D-17-0274.1