Review of "Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK" by Andres Peñuela, Christopher Hutton and Francesca Pianosi

This paper addresses comprehensively a topic that increasingly requires investigations: the value of seasonal forecasts for real-life applications. Until recently, many studies have worked on assessing or improving forecast skill, but still few manage to link this skill to value. In addition to being innovative, this study investigates the issue through different uncertainty lenses which makes it a very strong contribution to the field and results in valuable findings both for the scientific community but also for water management stakeholders. Overall, the paper, its ideas, structure and methodology are of high quality. However, additional information, for example about the data used and the forecast skill evaluation, would be necessary to support the analysis. In addition, I have some concerns about the validity of the results on linking skill and value due to the chosen methodology. These are detailed hereafter.

General comments

- The paper would benefit from a Data section, presenting for instance the data used as reference in the bias correction, the inflows used as reference in the forecast evaluation, or the demand model/observations.
- Some additional information would be needed in Section 2.2 on the forecast skill evaluation. More specifically, (1) In Figure 2, two inflows feed the two reservoirs (Is1, Is2), which inflow is being considered when evaluating forecasts? (2) Which time period is used to evaluate the forecasts? In Figure 8, November to April is shown, but in the forecast methodology (Sections 2.2 and 2.3.2) or in Figure 3, no specific time period is mentioned. I would recommend mentioning these two points in 2.1/2a.
- Since the goal of the paper is to assess the added value of dynamical seasonal forecasts for water management, and since authors evaluate the skill and performance of these forecasts against observations, it would be important: (1) to add information about how HBV was calibrated and setup for the area, or at least, to mention its performance (mean error) in simulating past inflows to the reservoirs (giving the possibility to make a parallel with the results in Figure 3); (2) to mention even briefly the hydrological regime upstream the reservoir system, as well as the interannual variability, which will define the added value of a method like DSP over ESP.
- I have concerns about the methodology chosen to link value and skill, and therefore about some of the subsequent results (first two paragraphs of Section 3.2.3). The authors are trying to link a skill obtained from comparing dynamic forecasts with forecasts based on past climatology (1981-20XX), with a value obtained from comparing dynamic forecasts with a worst-case scenario (1975/1976). On the one hand, the skill is based on the comparison of DSP-corr with ESP, meaning that its value solely indicates the gain in performance from using ECMWF SEAS5 instead of historical precipitation and temperature. On the other hand, the value is obtained by comparing DSP-corr with a benchmark based on a worst case scenario (1975-1976). These choices result in skills and values that cannot be compared. Instead, the skill computed in the paper could be related to the gains/losses in value between DSP-corr

and ESP (difference between the blue and green bars of Figure 7). Reversely, the value computed in the paper could be compared to a skill whose benchmark would be the performance obtained by using the worst case scenario (1975-76) as forecast in all years. This is a major point that should be addressed prior to publication to ensure the validity of the conclusions.

Specific comments

L45: “and reflects the risk-averse attitude”

L68: “to continue increasing in coming years”

L75-76: More literature review would be needed on past works on seasonal (climate or hydrological) forecasts value. Here are a few examples to consider:

Bruno Soares, Marta, Meaghan Daly, and Suraje Dessai. ‘Assessing the Value of Seasonal Climate Forecasts for Decision-Making’. *WIREs Climate Change* 9, no. 4 (2018): e523. doi:10.1002/wcc.523.

Giuliani, M., L. Crochemore, I. Pechlivanidis, and A. Castelletti. ‘From Skill to Value: Isolating the Influence of End-User Behaviour on Seasonal Forecast Assessment’. Hydrology and Earth System Sciences Discussions 2020 (2020): 1–20. doi:10.5194/hess-2019-659.

Parton, Kevin A., Jason Crean, and Peter Hayman. ‘The Value of Seasonal Climate Forecasts for Australian Agriculture’. *Agricultural Systems* 174 (2019): 1–10. doi:10.1016/j.agsy.2019.04.005.

L107: In this sentence, the authors seem to assume that their results could be generalized to extra-tropical regions and that the potential for use is region-dependent. However, as the authors state, results (forecast value) are in fact highly dependent on the investigated system, and therefore this sentence may sound a bit too ambitious compared to the paper objectives.

Section 2.1: I found this section very clear and helpful.

L118: Here, I was wondering what type of demand model was used. It is later in the discussion that the authors mentioned that the demand is based on observations. This would be worth mentioning earlier in the text, and would fit in a Data section.

L142 (2.a): More information would be needed at this stage about the forecast skill evaluation, for example by mentioning the reference and the benchmark used.

L160-161: The notation for units is not consistent between the text (Ml) and Figure 2 (ML).

L176: Did you mean a “reasonable chance”?

L182: Please consider adding the following reference for ECMWF SEAS5:

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H. and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12(3), 1087–1117, doi:10.5194/gmd-12-1087-2019, 2019.

L182-183: I suggest writing “The ECMWF SEAS5 hindcast dataset” because it is the hindcast dataset that includes 25 members, while the real-time operational ECMWF SEAS5 has more members.

L187: There are two points on this line.

L195: This statement needs to be refined. Linear scaling and distribution mapping may lead to similar results in terms of bias removal, but they will have very different results on the forecast themselves, e.g. linear scaling will just shift the ensemble, while distribution mapping will have a different correction for each member, resulting in larger impacts on the spread.

L196-198: Usually a difference approach is applied to calculate the correction factor for temperatures (see e.g. Lucatero et al. 2018, Teutschbein and Seibert 2012). Similarly the correction factor is added/subtracted to correct the raw forecasts. If a ratio approach was applied for temperatures, I would suggest adding a short justification of this choice. For instance, how are negative temperatures handled?

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J. and Jensen, K. H.: On the skill of raw and post-processed ensemble seasonal meteorological forecasts in Denmark, Hydrology and Earth System Sciences, 22(12), 6591–6609, doi:10.5194/hess-22-6591-2018, 2018.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456–457, 12–29, doi:10.1016/j.jhydrol.2012.05.052, 2012.

L206: “with what was done”

L208: The ensemble size has an impact on the CRPS (Ferro et al. 2008) and therefore, on the skill when the benchmark has a different ensemble size than the evaluated system. This could impact the results presented in this paper, and I would suggest adding comment about the impact of the varying ESP ensemble size on your skill results.

Ferro, C. A. T., Richardson, D. S. and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Met. Apps, 15(1), 19–24, doi:10.1002/met.45, 2008.

Section 2.3.5: I would suggest presenting the CRPS (L245-252) before presenting the CRPSS (L239-244) since it would define the CRPS notation before using it in the skill equation. There are also inconsistencies between the first two equations (CRPS and CRPSS) and the third one (mean error), such as the notation for the observations (*y* vs I^{Obs}), the notation for the system (*Sys* vs *Syst*), the mention of the lead time only in the third equation but not in the previous ones.

L237-238: This sentence is repeated twice.

L243: By choosing ESP as a benchmark, you also make the decision of only analysing the added skill from using dynamic weather forecasts over meteorological history, it would be worth mentioning.

L256: Do I understand correctly that the mean error is computed over a time window that varies with the lead time you consider, for example to evaluate lead time 6 weeks, you average the mean error from week 0, 1, … to 6? Is the same calculation method applied when computing the CRPS?

L279: I suggest replacing “gets lower” with “gets larger in absolute value”.

L334: The reference should be to Figure 5 instead of Figure 6.

L358: “two specific years”

L381-393: In this paragraph, conclusions are likely only valid for linear scaling, and not for any bias correction. I would suggest being more specific throughout the paragraph.

L446: "that should be kept in mind"

Figure 6: I wonder why, for each of the five operation scenarios, the authors do not display both DSP-corr and ESP in the same graphs.

Figure 6b and 6g: Shouldn't the absolute mean error be used to assess the correlation between the increase of resource availability and the performance, since both high positive and low negatives reflect poor performance?

Figure 7: It could be interesting to see the mean error and the CRPSS in this figure, even if they are not correlated.

Supplementary material – Figure 8: In the legend, I could not remember that 1975-1976 was the worst-case scenario. It could be worth mentioning it again at the end of the caption.

Supplementary material – Figures 9 to 13: The CRPSS is used to try and explain the gains in value from using ESP. Here, it is not clear which benchmark the performance of ESP was compared to in order to compute the skill, if it is indeed the CRPSS of ESP being shown. If not, then please refer to the fourth general comment.