Reply on RC1

Donghoon Lee et al.

Author comment on "Unfolding the relationship between seasonal forecast skill and value in hydropower production: a global analysis" by Donghoon Lee et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-518-AC1, 2022

Summary

The manuscript by Lee et al. explores the relationship between forecast skill and value in the case of the management of hydropower dams. The authors use dam characteristics and forecast skill to identify categories of dams that (1) show potential for improvement or not over climatology-based operating rules, and (2) show improvement or not based on realistic forecasts. A climate classification is further used to "regionalize" the added value of long-term forecasts for the hydropower sector and identify regions where improvements of currently low quality forecasts would translate into added value for dam management.

The paper is of very high quality, is well referenced, well written and scientifically sound. It will be undeniably valuable for the forecasting community, but also has potential to reach hydropower production managers. Along with the manuscript come supplemental materials that further detail the methodology and the results, as well as a dataset and an R script that allow readers to access the datasets for each dam.

For these reasons, I strongly recommend this manuscript for publication. Hereafter, I list some questions to the authors, some recommendations for improving explanations, and mostly minor points.

Response: We thank the reviewer for the positive comments and further critical comments that we believe will enhance the overall quality of the manuscript.

General comments

Sections 2.1 and 4.4: You decided to use the Köppen-Geiger climate classification. Since you are working hydropower and inflows which are influenced not only by climate patterns but also hydrological ones, a classification based on hydro-climate characteristics and not only climate characteristics would seem more relevant for the goal you are trying to achieve. Please consider using the hydro-climate classification proposed by Knoben et al. (2018).

Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, Water Resources Research, 54, 5088–5109,https://doi.org/10.1029/2018WR022913, 2018.

Response: Thanks for suggesting the HydroIogical Climate Classification (HCC) (Knoben et al., 2018). We agree with the suggestion, so we have investigated it and found the following two points. First, the HCC is derived from climate variables, such as precipitation, temperature, and potential evapotranspiration (CRU TS v3.23), and then is evaluated with independent streamflow data. Therefore, the HCC could still be seen as "climate-based" classification (although Knoben et al. (2018) showed that the HCC better represents streamflow characteristics in terms of grouping catchments). Second, the HCC is not a categorized classification like the Köppen-Geiger climate (KGC), but rather a set of three dimensional numerical climate indices, including annual aridity, aridity seasonality, and precipitation-as-snow. As a result, categorizing dams according to their hydroclimatic characteristics is not straightforward.

Although it is beyond the scope of our original goal--which is classifying forecast accuracy over climate zones--we analyzed the relationships between HCC indices and forecast skills of 735 dams (see the figure below). Here, we used averaged HCC values in the grids upstream of each dam. Figure 1.1 (reported below) illustrates forecast skills of 735 dams at the given HCC values. Some interesting patterns can be identified, such as dams in the snowing regions (snow ≥ 0.2) tend to have good forecasts with higher seasonality (seasonality ≥ 0.4) (panel b) or less aridity (aridity < 0.4) (panel c). We believe such interpretation can contribute to characterize dams' forecast skills at the given hydrological and climate conditions. However, we plan to include this analysis in the supplementary information according to the aforementioned two points.
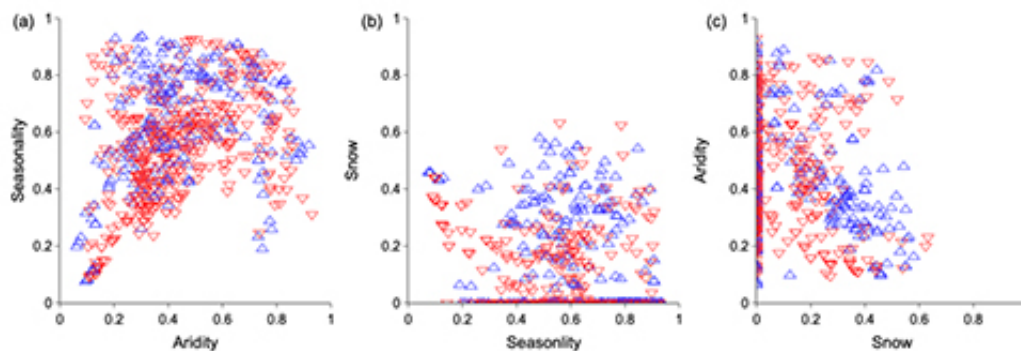

Figure 1.1. Scatter plots of HCC indices (Knoben et al., 2018) of 735 dams: (a) seasonality and aridity, (b) snow and seasonality, and (c) aridity and snow. Blue up-pointing (red down-pointing) triangles represent dams with good (poor) forecast skill based on X_{MdAPE} cuf-off value.

Section 3.1, Figure 1: It is not clear to me why the authors allow future inflows (t+1, t+2, ...t+7) to be predicted based on future climate indicators (1-8 months ahead). In a true forecasting setting, the ENSO, PDO, NAO and AMO indices for the 1-8 months ahead would not be available, only forecasts of these indices would. Some clarifications would be needed on this aspect. For instance, the authors could re-use the very clear notation t, t+1, …, t+8 to define which time steps they use in terms of climate teleconnection indices with respect to the forecast month t.

Response: We apologize for the misunderstanding. We used historical (t-8 to t-1) climate indices (ENSO, PDO, NAO and AMO.) For clarifying this, we will change Lines 148-150: "Then, we estimate the lag-correlations between future monthly inflows over the next 7 months (t+1 to t+7) and historical climate indices (t-1 to t-8), snowfall (t to t-8), and inflow and soil moisture in current month (t)." We will also update Figure 1.

Section 3.3.2: Even though the authors argue that MdAPE has a higher correlation and that it provides a value at each time step, KGE, and in particular its components, may

have given insights into the forecast characteristics (correct timing, volumes, variations) that influence value. This information would be extremely valuable to guide further model and forecast developments for the hydropower sector, in the same way your investigation of dam characteristics informs dam managers of potential forecast value. I wonder whether this would be something to explore also to address the limitation you note in the Results section L.341 "For dams with poor IDF and high KGE, two features are noteworthy: first, KGE may not fully capture the relationship between forecast skill and value".

Response: Thanks for pointing this out. We agree that KGE and its components may provide meaningful characteristics of forecast, hence we looked into the correlation between these forecast characteristics and the performance metric I (Table 1.1). The correlation drops with longer lead-times, which is expected, suggesting that better prediction for immediate months tends to lead to higher forecast value. However, this trend is not observed for the bias ratio (beta), which suggests that accurate prediction in inflow volumes for all seven future months contributes to higher forecast value. Yet, the correlation values here are still lower compared to other indicators of forecast skill presented in Table S4 (e.g., correlation between I and MdAPE is -0.4). We think including these results in the supplemental material may therefore be the best option.

	MP1	MP2	MP3	MP4	MP5	MP6	MP7
KGE	0.21	0.15	0.14	0.11	0.09	0.07	0.06
r	0.23	0.16	0.15	0.12	0.10	0.08	0.06
beta	0.17	0.20	0.16	0.16	0.19	0.20	0.17
gamma	0.17	0.11	0.12	0.08	0.08	0.06	0.04

Table 1.1. Correlation between performance metric I and forecast skill for the 269 dams that are classified as success cases. Forecast skill is represented by KGE and its three components, r (correlation), beta (bias ratio of mean inflow), and gamma (variability ratio). The columns correspond to the prediction model with 1 to 7 months lead-time.

Forecasts with horizons up to 7 months are generally probabilistic to account for uncertainties at such long lead times. The authors should discuss the role of uncertainties in their study design, i.e. how realistic it is to consider the value of deterministic long-range forecasts depending on the current state of hydro-climate long-range forecasts, but also on the capacity for hydropower dam managers (whose actions are hypothesized in this study) to inform their decisions based on probabilistic information.

Response: Thanks for pointing this out. We agree this is a point worth discussing. In particular, we plan to do so at Lines 464-465. The text reported below provides an anticipation of the discussion we plan to include.

"Finally, investigation of alternative forecast approaches may be warranted. The adoption of a statistical prediction model is motivated by the availability of relatively long hindcast periods and the need for long prediction horizons (Section 3.1). Particularly, a probabilistic forecast, which is typically based on a statistical model with empirical distributions or ensemble dynamical forecasts, could be used in place of our deterministic forecast to reflect a more realistic dam operation, such as decreased dam efficiency over a longer lead-time due to increased forecast uncertainty. Additionally, depending on the reservoir specifications, probabilistic forecasting offers a greater potential for improving dam operation than deterministic forecasting (Zhao et al., 2011). Thus, incorporating probabilistic forecasting into the design of our study will allow for a more accurate quantification of the realistic forecast value and the impact of prediction uncertainty in

relation to reservoir characteristics."

Specific comments

L.132-135 There is a range of models that fall between statistical prediction models and physically-based model. For instance, conceptual models do not fit in these two broad categories. I invite the authors to revise this statement.

Response: We agree with this point. For clarifying that, Lines 132-135 will be changed to:

"Seasonal streamflow forecasting approaches in large-scale analysis include physically-based (mechanistic) models, such as GloFAS (a global-scale forecasting system; Emerton et al. (2018); Harrigan et al. (2020)), empirical or statistical (data-based) models that leverage the relationship between large-scale climate drivers and local hydro-meteorological processes (Block, 2011; Gelati et al., 2014; Giuliani et al., 2019), and conceptual (parametric) models that integrate hydrological processes at the catchment scale (Lindström et al., 2010; Devia et al., 2015)."

L.137-140 The arguments for choosing a statistical model rather than a physically-based one seem too general. In fact, the statement "the prediction horizon of most physically-based approaches (a few days to 3-4 months) falls short of our preferred lead times up to seven months" only holds when considering currently openly available global reforecasts, and not reforecasts from physically-based (or rainfall-runoff) models in general. There already exists, for instance, global reforecasts up to 7 months ahead and with hindcast periods for at least 30 years (https://hypeweb.smhi.se/explore-water/forecasts/seasonal-forecasts-global/). As the authors rightfully mention in the section on opportunities, "global-scale forecasting systems are gaining momentum", and therefore this part should be rewritten to highlight the impermanence of the statements.

Response: We agree with this suggestion. To reflect it in the manuscript, we plan to update Lines 136-141 as follows:

"Two broad alternative approaches for seasonal streamflow forecast development include physically-based models, such as GloFAS (a global-scale forecasting system; Emerton et al. (2018); Harrigan et al. (2020)), or statistical prediction models that leverage the relationship between large-scale climate drivers and local hydro-meteorological processes (Block, 2011; Gelati et al., 2014; Giuliani et al., 2019). Here, we select the second approach for two reasons. First, the prediction horizon of most currently openly available global reforecasts (a few days to 3-4 months) falls short of our preferred lead times up to seven months, needed to test the potential of realistic forecasts for a broad spectrum of reservoirs—including those characterized by slow storage dynamics. Second, re-forecasts issued by global-scale forecasting systems are only available for a relatively-short hindcast period (typically two decades; Harrigan et al. (2020)), whereas the time series of globally-available hydro-climatological data are significantly longer. It should be noted that these two statements may change in the near future as the boundaries of global-scale forecasting systems keep getting extended (see Section 5.2). For example, there already exists global re-forecasts from physically-based models with a prediction horizon of seven months and hindcast periods of about 30 years (https://hypeweb.smhi.se/explore-water/forecasts/seasonal-forecasts-global/)."

L.141 "Our long-range inflow prediction **model** uses…"

L.145 "For example, forecasts **issued**…"

Response: Thanks for spotting these two typos. We will correct them.

L.174 Isn't it the goal of the dam inflow prediction model to feed the reservoir model? If so, isn't Qt not only retrieved from WaterGAP but also from the proposed statistical dam inflow model?

Response: At month t, the prediction model gives estimates of Q_t to Q_{t+6}, which consequently determine the release sequence R_t to R_{t+6} (in Eq. 4). Then, only the decision R_t is implemented. When simulating the reservoir dynamics, the observed inflow Q_t retrieved from WaterGAP is used.

Section 3.2.1 Is there any need for initialization of this reservoir model, and if so, how do you handle this aspect? e.g. which initial values do you use for instance for the reservoir storage?

Response: All reservoirs begin at full storage at the start of the simulation period (i.e., 1958). We will clarify this point in the revised manuscript.

L.227 "… may influence ..."
L.234 "It is reasonable to hypothesize that the value…"

Response: Thanks for spotting these typos. We will correct them.

L.251-253 "Note that failure implies that the control rules and perfect forecast-informed operations generate a similar amount of hydropower, meaning that information on storage and previous-month inflow are sufficient for near-optimal release decisions." Wouldn't that correspond to an IPF value of 0 rather than to the mean IPF? If this statement is based on the mean IPF value, the reader does not have this information yet, and this sentence is confusing.

Response: We agree that this statement is confusing, since 'failure' in this case actually means that the reservoir has IPF value < mean IPF which could be (and in most cases, is) greater than 0. We thus plan to change the term from "success/failure" to "case/non-case" and remove lines 251-252 to avoid the confusion. The text will be thus changed as follows: "First, for each dam, we label it as case (also referred to as success) if it has the desired property of a IPF value larger than the mean value of IPF across all dams. Otherwise, the dam is labeled as non-case."

L.320-321 "Considering the superior performance of the MP1 model, the forecast skill of MP1 only is retained to represent the overall forecast skill in the following analyses." Since the optimisation uses all forecast horizons, the speed with which skill decreases with the forecast horizon may play a role in the optimization and could have been considered as well.

Response: Thank you for raising this point. It is indeed true that the speed with which forecast skill decreases may play a role. To investigate this, we first fit a linear regression between KGE and prediction lead time for each of the 269 dams classified as success cases. We then use the slope of the regression to represent the speed with which forecast skill decreases (i.e., a highly negative slope means forecast skill drops quickly with longer lead-times). We then plot the performance metric I against the slope (Figure 1.2). There is no clear trend of correlation between the speed with which forecast skill decreases and forecast value. Given this result, we believe that such analysis should be reported in the supplemental material, but not in the main manuscript.
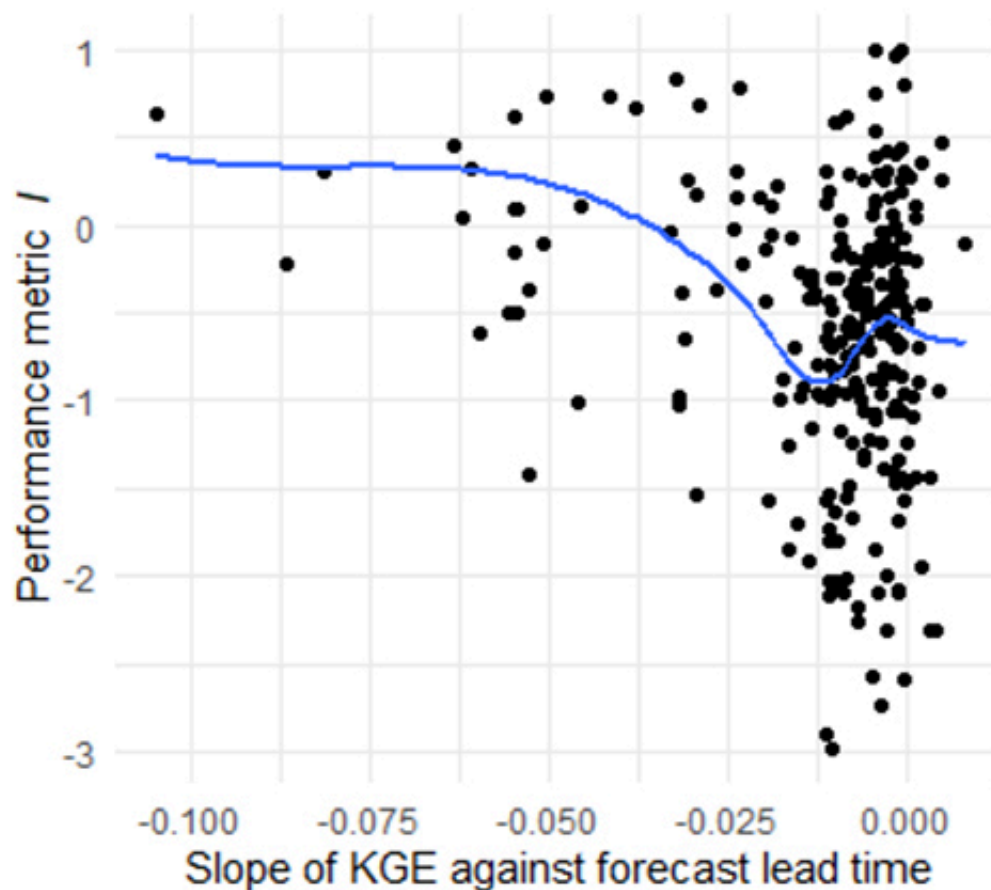
Figure 1.2. Scatter plot of performance metric I and slope of KGE against forecast lead time for the 269 dams classified as success cases. The blue line represents the local polynomial regression fitting performed on the data points (i.e. fit at point x is done using points in the neighborhood of point x).

L.327-329 "Small negative values of IPF are likely a result of the discretization needed by dynamic programming to optimize the release sequence (eq. (4)), hence allowing control rules to outperform perfect forecast-informed operations." Could you please further explain what you mean to help understand the counter-intuitive negative IPF values?

Response: Release decisions are discretized into 20 levels while storage is discretized into 500 levels. Hence, when storage level falls between 2 discrete levels, the closer level is selected and the optimum release decision for that discrete level is implemented. This decision may hence be suboptimal sometimes, giving rise to the negative IPF values, which are nevertheless small in absolute terms and do not affect the interpretation of our results. We will clarify this point in the revised version of the manuscript.

L355-357 "This is attributed to the weekly operations, suggesting that more frequent release decisions may reduce forecast value, since the benchmark operating rules have more opportunities to adjust release decisions." Isn't it the case for all the dams below the horizontal line? Why should these ones (the failing ones) behave differently?

Response: Our original intention was to make a comparison between weekly operations and monthly operations. For all dams below the horizontal line, weekly operations reduce forecast value as shown in the figure below. We did not mean to make a comparison between the failing dams and the successful ones below the horizontal line. We realize that this statement can be misleading, hence we will revise it to clarify our point:

"This is because weekly operations decrease IPF for some of these dams to below the mean IPF, turning them from cases (if operated on a monthly basis) into non-cases. This suggests that more frequent release decisions may reduce forecast value, since the benchmark operating rules have more opportunities to adjust release decisions."
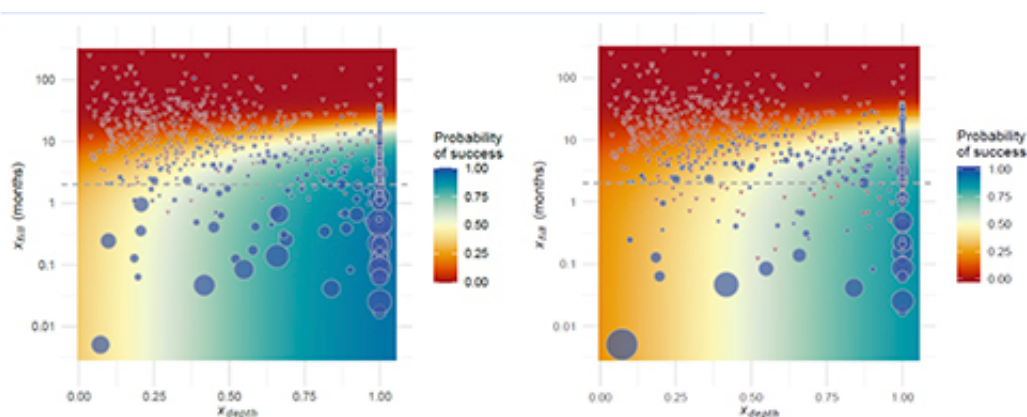

Figure 1.3. Probability of success estimated using logistic regression when all dams adopt monthly operations (left) and when smaller dams (below the dashed line) adopt weekly operations (right, same figure as in Figure 5). Weekly operations tend to reduce forecast value as shown by the smaller blue circles and greater number of non-cases (triangles) in the right figure.

L.446 "… a number of assumptions that must be properly contextualized."

Response: Thanks for spotting this.

Figure 2 It would be more correct to change the caption to "Percentage of dams whose inflow is significantly correlated with…" since a dam in itself is not correlated to anything.

Response: We will revise the sentence as the reviewer suggested.

Figures 3 and S2 The titles for the color scales in the second and third lines of this figure are confusing. If my understanding is correct, I would suggest changing titles in the first column to "Change in number of predicted months", and in the second column "Change in KGE", with "(b) MP4-MP1" and "(c) MP7-MP1" on the left-hand side.

Response: Thanks for your suggestion. We will modify both figures as suggested (please refer to the figures reported below).

Figure 5 "red triangles represent failures"
Figure 6 "meaning that the performance of realistic forecasts is worse than the one attained by control rules"

Response: Thanks for spotting these typos. We will correct them.

References

Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, Water Resources Research, 54, 5088–5109, https://doi.org/10.1029/2018WR022913, 2018.

Zhao, Tongtiegang, Ximing Cai, and Dawen Yang: Effect of streamflow forecast uncertainty on real-time reservoir operation, Advances in Water Resources 34.4 (2011): 495-504.