Thank you for this thorough and very complete review. I apologize for the lack of precision in our terminology and any impact it might have had on your understanding of this study. We still hope there is some time for discussion if there is additional interest.

Yes, it is true that Hydrometeorological prediction was not our main research interest at the start of our project, although some of our co-authors and project partners are highly experienced scientists in weather forecasts.

We will make sure to use the standard terminology used in this field. Thank you for picking those multiple loose threads that needed further attention. We will make sure the terminology and writing style is unified throughout the manuscript.

Below are some preliminary responses to specific points the reviewer raised ("excerpts of the reviewer's comments are underlined for clarity").

windows of opportunity: these windows of opportunity are indeed not properly defined in the text. Mentioned in the introduction l. 89-93. A window of opportunity will be defined as a skillful forecast for a given variable's tercile during a given season. In other words, a trustful forecast that a stakeholder can use as a decision support information. For example, the lower tercile of surface temperature in Spring at our Norwegian site is a window of opportunity (Table 4). In other words, our workflow is able to forecast when surface temperature will be lower than normal in Spring at the Norwegian site. We focus on those because the whole project and study here is to be seen from a stakeholder point of view, and we would like to focus only on forecasts that we are confident are the most trustful. This point will be made clearer from the start, to also respond to your comment on l. 194-197.

Hindcasts: Sorry for the confusion here. Yes we used hindcasts for the 92 three-month seasons (11/1993 to 11/2016) (see L. 170). We refer to 1994-2016 for simplicity, but we will make sure that everything is clear and consistent throughout the manuscript, also when the hindcasts were produced.

Another point is the source of predictability. We agree that the SEAS5 forecasting system cannot be seen as a source of predictability, it likely further transfers predictability from e.g., ENSO and NAO, as you mentioned, to our catchment and lake model. We will consider this point very carefully as it impacts the title as well. Thank you for picking this up. We will make sure to avoid these confusions throughout the manuscript (e.g., L. 53-55; L. 76-77 ...) and any other confusion caused by unprecise terminology related to forecasting (e.g., l. 167-168; l. 173-174; L. 174-175; l. 191; l. 347-348; l. 366-368). We will avoid any cryptic formulation.

SEAS5 system and ERA5 resolution and downscaling issues: We have used standard tools and have colleagues in the group that are experienced with this type of issues, so I know we have bias-corrected, down-scaled ERA5 and SEAS5 data in an acceptable way. We will provide further detail on these pre-processing steps in the revised version of the manuscript. For now, you can have a look at Mercado-Bettin et al. (2021; https://doi.org/10.1016/j.watres.2021.117286)

I'm really grateful to the reviewer for the quality and completeness of this review, Thank you for these specific and relevant comments. I'm trying to provide preliminary (for now) answers to some of the comments below:

More specific comments

Introduction

Line 24-26: Do you mean the skill of the meteorological predictions (SEAS5 outputs) are worse then the skill of the hydrological predictions?

> Yes this is what we mean here. Meteorological predictions were worse than the skill of the discharge (to only some extent) and lake water temperature predictions.

Line 70 "...temperature predictions and forecasts" and L. 72

Yes thank you, we will make sure to have a consistent terminology

Line 80-84: I do not understand what the authors try to argue, what is meant by "water flow predictability", do you mean discharge of the rivers? Lake level heights? Can you make this sentence clearer?

Yes we meant river discharge. We will make that clearer.

Line 90. We will introduce ice-off.

Line 90-91 and Table 4 (showing actual values of ROCSS and FRPSS) and Table 5

The aim of this table (Table 4) was rather to quickly show where and when the forecasts were skillful, and compare the skills of SEAS5 forecasts (climate predictions) with lake forecasts. We can provide all values in the supplementary to avoid overloading the table. Note that we haven't considered all forecasts with ROCSS and FRPSS higher than 0 as "skillful". We have described this L. 194-197:
"Threshold RPSS and ROCSS values above which RPSS and ROCSS are significant at 95% confidence are calculated by built-in VisualizeR functions and were used to identify windows of opportunity (i.e., combinations of seasons, variables and terciles for which forecast performance was significantly better than the reference). In our case, these thresholds typically range between 0.47 and 0.55."

Admittedly, our formulation is lacking clarity. We will provide more details in the revised MS in the text and the table caption. This is also linked to your comment on L. 262 on how fair we determine a RPSS is significant and Table 5.

Methods:

Line 101 to 111: Thank you for your nice suggestions. We will incorporate those.

Line 114-115: As stated above, we will provide more detail on the bias-correction method used (also from L. 125).

Line 116: term "impact models" and "impact variables".

Yes these are the water quality models and variables. We will unify these terms and them throughout the manuscript.

Line: 125

Unfortunately, we did not do any forecast verification with different bias-correction. Given the frame of the project and the resources it takes, it will unfortunately not be possible. On the other hand, we can look at additional scores and measures.

Line 127-128:

We apologize for the messy abbreviations of variables here, we will make sure those are defined and consistent throughout the manuscript. Given that we applied different models at the four sites, sometimes the forcing weather variable were slightly different. We will make sure to described that and harmonize everything.

Line 130-131: Yes we can provide further detail on the observations, these were collected by our institutes, or published datasets.

Line 143-145:

All hydrological models were calibrate against local observations, this is described in Mercado-Bettin et al. (2021; https://doi.org/10.1016/j.watres.2021.117286) but will be briefly described here again for clarity. Note that models were calibrated and validated against two different time periods following best-practices.

Line 152: Thank you, we will do.

158-159: "Most common statistical goodness-of-fit parameters, e.g., Kling-Gupta efficiency (KGE), NSE and RMSE, for hydrological and lake modeling were calculated."

We will explicitly describe the values of these performance criteria.

Line 174-175: Aggregation prior to forecast verification

We don't do any lead-time dependent verification, we only aggregated the forecast to seasonal means (1 value for the whole 3-month period for each year between 1993 and 2016) and used that in our verifications and calculations of skill scores. We will make sure to clarify this point in the manuscript.

Table2 and full paragraph:

Thank you for your suggestions here. Yes pseudo-observations were used as the reference forecast. In addition, we used also observations as the reference forecast, when and where it was possible (when data gaps where below a given threshold). We apologize again, this threshold was not described properly. Note however, that the "Obs coverage" in table 5 highlight the cases for which there are enough observations. This point will clarified.

Line 200-227: I struggle a lot to follow the explanation. First of all what is ROCSS**s/** ROCSS**w/** ROCSS**w+t**

Thank you for your suggestions on rephrasing the paragraph, confused formulations and updating the paragraphs' titles. These ROCSS are the skill scores calculated from the forecasts of the various sensitivity analyses (SA) described just before, where forcing data over the target season (S), the warm-up period (W) and warm-up plus transition periods (W+T) was replaced by random data. As described l. 216-217: "The outputs of S-SA, W-SA, and W+T-SA were used to produce tercile plots and calculate ROCSS. The comparison of the ROCSS values (ROCSS_i) obtained for the various SAs" We will clarify this formulation to make sure this point is not confused.

Line 240-244:

This will be clarified. In fact, we will provide lake heat budgets for each site to support this, as it was also raised by the other reviewer. This is supported by our data analysis and also shown in the literature (e.g., Blottiere, 2015).

Results

Line 251: We will add part of Table S2 into the main text.

Line 262: Sure, we will provide more background on fair RPSS and what it accounts for.

Line 269-270:

"Only 0 to 10% of the SEAS5 climate 270 hindcasts are skillful, on average".

I struggle understanding this result. The RPSS (or any score) is usually determine based on a large sample forecasts (or hindcasts). Of course, for an individual forecast this might be poor but the full picture of forecast performance can only be revealed when the Scores for many issued forecasts are investigated.

Here RPSS and ROCSS are determined based on the 23 seasonal means (spring, summer, autumn, winter) from 1993 to 2016. As described l. 189-191, Briefly, the RPSS provides a relative performance measure on how well the probabilistic ensemble is distributed over the lower, middle and upper terciles, while the ROCSS provides a relative measure of discriminative skill for each category.

I think here you should add a figure with the results where the reader can see how the hindcast performance actually is. From the text and the table alone, it is rather difficult to follow your argumentation. In addition, what is the definition of a skillful hindcast in your context? Is every hindcast with a ROCSS>0.5 skillful or do you use other thresholds? It would be crucial to mention the numbers at least once in the results section as well.

We were struggling to find a clear and concise way to create a figure showing hindcast performance but we were looking for this. So we would be very happy if you have a suggestion? Maybe a figure from a published paper than we can be inspired from. How would you like to see this information plotted?

Regarding skillful forecasts, we will clarify this. We basically consider that all hindcast with significant ROCSS are skillful forecasts (where the threshold is in between 0.47 and 0.55 (as described above).

Line 278-280: How many seasons are discarded due to missing observations? Please indicate the exact numbers such that the reader knows how many samples (hindcasts) are actually used for the analysis. Or at least refere to the table where the numbers are listed.

We calculated ROCSS_obs only when >50% of the seasons (i.e., 12 seasons) were represented by some data. But in practice, most of the variables for which we calculated ROCSS_obs had 100% (i.e., 23 seasons) covered by observations. See also “Obs coverage” in Table 5. We will introduced this notion of observation coverage earlier in the manuscript and make sure everything is clear how we calculated ROCSS_obs, including the number of seasons.

Line 283-286: Yes we will rephrase here for clarity.

Table 4

Again general comments: abbreviations are non intuitive. FRPSS is not introduced before. I do not get the message of this table. I would prefer to see the skill scores (e.g. as boxplots) over different seasons for the variables. It is not clear to me what temporal aggregation is the baseline of this analysis.

The main message here is to show that there is limited skill of the SEAS5 forecasts (climate predictions) but still some skill in the lake forecasts. In many cases, those are not synchronous, e.g., in Summer in Spain, 5 variables’ tercile are associated with significant ROCSS, what we call “windows of opportunity”, whereas there are only 2 weather climate variables’ terciles with significant ROCSS. Looking at this table, we were hoping that the reader would say, “Oh, there is forecasting skill coming from somewhere else that SEAS5, e.g., inertia”

Regarding temporal aggregation, this is again based on seasonal means. We apologize for the oversight, and will include that in the caption.

Line 305:

Fig. 31 should be Fig. 3 This is Fig. 3l with “l” for “Luke Skywalker” for panel “l”. We will use capitals to avoid misunderstanding.

Can you elaborate how the ROCSS is determined, what exact values are taken into account? Do you use daily values to calculate the scores or weekly/seasonal aggregated values? This is still unclear after reading the manuscript.

ROCSS are determined from seasonal means. We will make that point clear throughout the manuscript.

Fig 3: it is confusing that in the plot description and the text you mention ROCSSs etc. but on the x axis S-SA W-Sa etc are displayed. I suggest to unify all and make the plot more readable.

We will replace “S-SA”, “W-SA” and “W+T-SA” with their respective ROCSS_i expressions.

Line 320-322:

Maybe because I do not understand what the windows of opportunity are, I do not get the message here. I suggest to more explicitly formulate what the impact is of changing the initial conditions and the forecast input. The result indicates that it is not worth using seasonal forecasts at all, which is hard to believe. Please elaborate such that the message gets clearer.

We will describe better what is the impact of changing the initial and boundary conditions. We suggest that there is a lot of skill that originates from legacy effects from the catchment model, and form inertia of the lake/reservoir systems themselves. Only at the norwegian site for specific variables and seasons, there are signs that using seasonal weather predictions provide further skill. For the other sites, we cannot see any sign of this. So, yes for most of the forecasts that were significantly skillful, it was not worth using seasonal forecasts to force the models. We will make this point clearer.

Line 323: again the title does not reveal to me what will be considered in this paragraph. I suggest to use a more appropriate title for this paragraph.

Thank you, we will provide a more relevant title.

Figure 4: It is hard to follow what is shown here. What is the relative sensitivity. Maybe it helps if you refer to the exact paragraph number in the label of the figure. In addition, is there a reason why you use a color coding and a size coding? I suggest either using color or size, otherwise it seems that multiple aspects are coded.

Yes, we will refer to the exact paragraph number for the description of how relative sensitivity is estimated, see L. 229-234, section 2.1.1. And you are right, color and size show the same information. We will keep only the size coding.

Line 347-348: "Hence, a significant fraction of predictability is originating from the SEAS5 dataset although the largest source remains ERA5 data over the warm-up". This sentence illustrates what I think makes the manuscript complicated to follow. The reader himself must make the connection what this means. If I am correct, it shows that the initial conditions are more important than the driving meteorological predictions. Is this correct? It would make the manuscript much more readable if you directly refer to formulations what it actually means in addition to just give such "cryptic" explanations.

Yes you understood this sentence right, and again we apologize for the cryptic formulations. Learning the forecasting terminology as we go. We will improve that.

Line 394-395:

Literature on streamflow hindcasts broadly shows that beyond the transition month, climatology-driven hindcasts are typically 395 more skillful than hindcasts driven by seasonal climate predictions (Arnal et al., 2018; Bazile et al., 2017; Greuell et al., 2019).

This is a misleading interpretation, when you say "climatology driven hindcasts". In all three publications a well established ESP (ensemble streamflow prediction) approach is used. This can be seen as a climatology driven hindcast, but for a scientific publication I would expect to have a clearer formulation. In addition, in these papers there is no transition month mentioned, a concept I do not understand. All papers mention the first lead time month. What exactly is the transition month in your analysis?

We will be more precise when referring to these studies. The main point here is that hindcasts driven by seasonal climate predictions are not necessarily more skillful.

We will replace the “transition month” with lead month 0, in agreement with Greuell et al., 2019, i.e., the month following the date on which the forecast would have been issued. If the forecast was issued on February 1st, 2016, everything before that date is part of the warming period, February 2016 is what we used to call the “transition month”, lead month 0, and the target season is the lead month 1 to 3: spring 2016, march to May 2016. This clarification will be included in the method section.

Again, thank you for this thorough review and for giving us the opportunity to improve by kindly pointing to our lack of precision in our terminology. Most of reviewers would just not bother to provide such constructive feedback.