Dear Editor,

We would like to thank you again for the opportunity to revise and resubmit our manuscript with the title "The suitability of a hybrid framework including data driven approaches for hydrological forecasting" (manuscript number: hess-2022-89) after the first round of reviews. Furthermore, we would like to thank the two reviewers for their continuous remarks and suggestions, which were helpful in improving the manuscript.

We have carefully considered and responded to each comment (see response to the reviewers below) and made changes to the major issues mentioned by the reviewers. We have provided additional details and made modifications:

- We added additional explanations on our modelling framework (section 2.1.2 and appendix A) suggested by both reviewers
- We added clarifications throughout the manuscript regarding the terminology surface water levels (main comment reviewer 1)
- We rephrased parts and added additional information for different result sections as well as extended the discussion (section 3.2, 3.3 and 4) suggested by the reviewers
- We updated figure 3 and B2 for clarifications

In the author response copies below a point-by-point treatment of the reviewer comments is included. We repeat the comments of the reviewers in italics and our response in roman. We refer to the lines in the revised paper to denote the changes implemented and corresponding changes are highlighted in the manuscript text in the revised file.

Thank you again for your consideration of our revised manuscript.

Sincerely,
Sandra Hauswirth, on behalf of the coauthors

Anonymous Referee #1

Referee comment on "The suitability of a hybrid framework including data driven approaches for hydrological forecasting" by Sandra M. Hauswirth et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-89-RC1, 2022

General comment:
The authors present a hybrid forecasting framework combining data driven approaches (using local, in-situ observations) and seasonal reforecasting information from large scale models to predict hydrological variables. The authors show that skillful predictions can be obtained with this hybrid framework. Although the idea of this framework is innovative and deserves publication a major revision is required (see comments below).

Response: We want to thank the reviewer for taking the time to review our manuscript and acknowledging the innovative aspects of our research. We will address the comments, suggestions and open questions point by point below.

Major comment:
As suggested in the title and also throughout the manuscript a heavy focus for the assessment of this hybrid framework is based on the prediction of the variables river discharge and surface water levels, which seems to refer mostly to river water levels (it is actually not specified whether surface water levels refer to river water levels, sea level or even lake levels). However in the introduction and in section 2.2 the usage of sea water levels is mentioned. Furthermore, Fig A2 as well as Fig. 3 suggest indeed that sea water level observations are being considered. In the remainder of the manuscript the authors do not distinguish between sea level and river water levels but only mention surface water level measurements and it seems that in some of the analysis water level measurements from rivers as well as sea levels are mixed together (e.g. Fig. 4, Fig. A1, Fig. A3, Fig. A4). The mixed results are then used to derive general conclusions about the predictive skill of the hybrid framework. For example, in line 202 it is mentioned that Fig. A1 representing the CRPSS for surface water levels shows even better performance than the one for river discharge (Fig. 2). If sea and river water levels have been merged together it is however not possible to do such a comparison as the underlying processes that drive changes in sea level and river water levels are different. In addition to the mixing of those two variables, conclusions are made throughout the manuscript which are mostly only applicable to river water levels and river discharge. For example, lines 225-243 describe the results for the station Hagestein Boven and state that the increase in skill in the early spring months are due to the fact of snow melt dynamics. Obviously this conclusion is not valid for the results obtained from sea level stations. However, no further analysis is provided for the skill observed in sea level stations. The same is true for the section 3.2 on hydrological low flows which also is not applicable to sea level measurements. Instead, section 3.3 mentions surface water level predictions (and it is not clear whether this refers only to river levels or to sea levels or to both) and makes some general conclusions but does not provide any further detail. Even in the introduction the manuscript provides primarily references in relation to streamflow forecasting and fresh water management but does not make any reference to coastal water level predictions. In my view, the authors have two options to improve this issue: 1.) either you focus your analysis only on fresh water, i.e. only on river discharge and river water level predictions and remove from the analysis all sea level predictions or 2.) the authors clearly separate the results and their analysis for sea level predictions from river discharge/water levels predictions expanding the manuscript with the relevant sections and presenting separate conclusions/discussions for sea level and river flow/level predictions.

Response: We appreciate the reviewer's input and remarks on the interpretation of our results and the combination of our datasets included. However, we believe that there is a misunderstanding regarding the model framework and its set up, which is leading to the misunderstanding in the results interpretation.
The modelling framework is based on a previous study by the main author and we acknowledge that the description in the current manuscript might not have been sufficient to fully follow without consulting the previous study. Because of this, the term "surface water levels" appears not to be properly coined in the current manuscript. This, in turn, may have led to the issue raised by the reviewer that surface water levels not being clearly enough defined and leading to an erroneous interpretation of the results. When we refer to "surface water levels" in our paper we mean water levels of <u>rivers, streams and lakes</u> Thus, we focus only on fresh water flows and fresh water levels as forecasting targets.

The results for the different locations (as shown in Fig A2 and A3) thus pertain to forecasts of fresh surface water levels that are based on the machine learning models trained and validated for that specific location. As input data for the machine learning model the following variables were considered: discharge of the two main rivers entering the Netherlands, precipitation and evapotranspiration of one meteorological station in the centre of the Netherlands, as well as sea level observations close to one of the major dam systems at the coast of the Netherlands. Thus, sea water levels are an input variable to our machine learning based predictions and are not predicted themselves. When training of the machine learning models, observation of both input variables (discharge of the two main rivers entering the Netherlands, precipitation and evapotranspiration of one meteorological station in the centre of the Netherlands) and output variables (river discharge and river water levels) were used. When forecasting the trained machine learning models were forced with forecasts of the input variables. So all conclusions and interpretations pertain to forecasted fresh (river, streams) surface water levels

We will pay special attention to clarify this during the revision of our manuscript to prevent future confusion. This would also be in line of the reviewers first suggestion. The clarifications in the revised manuscript can be found in line 82, 102, and 142 in the extended model description of the track change file.

Other comments:
Introduction: Whereas the introduction mentions various examples for streamflow predictions no example is mentioned for sea level/coastal predictions that would support the integration of sea water levels into this analysis
Response: In line with the previous explanation we will revisit the introduction and adjust it to make it more comprehensible

Materials and Methods: Please add a section that describes the number of observation stations, its locations, its observation record, the variables used (river discharge, river water level, sea level) that have been used in the manuscript for training the ML models and that have been used for the analysis. Figs 3 and 4, and Figs. A2 and A3 show different station locations and it is totally unclear which observations have been used in this manuscript.
Response: In addition to the current link to the previous study where the modelling framework was developed, trained and tested, we plan to expand the explanation on the modelling framework so that it will also be followable for this study. We added an extended model description in section 2.1.2 and included an additional figure to highlight different station locations and input variables in the appendix of the track change file.

Figures 3 and A2 are not readable! Please increase the legends!
Response: We will revisit the figure and improve the readability

Figure 4 shows a station along the coast but is showing the ACC for discharge hindcasts. How is that possible? Or is there actually a small river, which is not shown in the Figure, flowing into the sea for that station?
Response: The reviewer did see this correctly. The current figure only shows the main river network, however there are many smaller rivers and streams that are unfortunately not depicted. The station shown close to the coast is a measuring site placed at one of the sluices, which is also connected by to the main rivers by a smaller river branch. We will make this information more clear in the manuscript. We added an additional comment in figure 4 and B2 of the track change file.

Section 3.2: It is stated that "BSS confirms the earlier findings and shows the same trend of increased performance in the first lead weeks….". I disagree with this finding as Fig. 6 shows clearly that the BSS for Feb/May and Sept is low in contrast to the findings for the general performance.
Response: We do acknowledge and explain the low performance in the following sentences: "The BSS confirms the earlier findings and shows the same trend of increased performance in the first lead weeks, additional skill of several weeks is found for early spring and early summer months. However, tiles with lower performance throughout long lead periods, late summer and winter months can be spotted for this station. Some of these weeks appear to be more difficult to predict compared to early months in the year. This is likely due to unequal distribution of low flow occurrences throughout the year: where during summer months low flows can be more common and therefore chances to not fully capturing every event are higher, the low flows during winter are less common and captured relatively well with the snow melt dynamic as seen in previous scores."

We acknowledge that the first sentence might be misleading and we will adapt it. We revisited the section 3.2 and rephrased our description of the results.

Section 2.2.: Please add a very brief explanation to the lagged times series approach as most of the readers will not be familiar with this approach.
Response: We will include a brief explanation of the lagged time series approach together with the extended explanation of the modelling framework. However, we would also like to highlight the reference to the previous paper, in which the development and further detail of the modelling framework are shown. We added additional information in line 125 of the track change file.

Lines 205-212: It is stated that only minor differences are observed for the different ML models. Please analyse better the reason for this. One would assume that advanced DL methods such as LSTM would perform better than multiple linear regression
Response: This is indeed a remarkable result, since when used with observed input data, the differences between the different ML models were larger. The most likely explanation is that the forecasting skill is very much dependent on the skill by which the input variables are forecasted, which apparently make the differences in skill between the ML models insignificant in comparison. We will add an explanation for this in the manuscript. We added additional explanations in line 240 and 324-330 of the track change file.

Figure A1: Does this figure show combined results of sea level and river levels? If yes, please separate these two variables.
Response: Figure A1 only shows results of forecasted fresh (river, streams, lakes) surface water levels

Anonymous Referee #2

Referee comment on "The suitability of a hybrid framework including data driven approaches for hydrological forecasting" by Sandra M. Hauswirth et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-89-RC2, 2022

The study is building heavily on results published elsewhere (Hauswirth et al., 2021). I guess that is fine and always a difficult call to decide just how much information to provide so that a paper becomes a stand-alone piece of work without unnecessary repetition. However, in places I would have liked a little more info in this paper so I would not necessarily have to read the previous paper. For example, lines 114-117; this seems important and a bit more in-depth description of the datasets (SEAS5) and the 'lagged time series approach' which I am not familiar with.

Perhaps I have overlooked something, but I am not sure how the all the forecasts made by the different models and different ensembles are aggregated into one CDF (as per Fig.2)? Also, If there is only a minor difference in the performance between the five different ML models (line 205) then what is the advantage of using all of them rather than selecting the 'best' or most credible model for a particular site and use that? Might it be useful to add a section highlighting the performance of each of the five ML models in contrast to the aggregate performance of the entire system?

Response: We want to thank the reviewer for taking the time to review our manuscript and sharing thoughts and suggestions for improvement. We will address the comments, suggestions and open questions point by point below.
We agree that our study is heavily building up on a previous study by the main author, which was a conscious decision that was made. The idea was and is to use the base modelling framework for different experiments. Therefore, we wanted to fully focus on the model development aspect in the first publication and refer to it for future work. We acknowledge that in this case, the explanation of the modelling framework was too short (as also indicated by the other reviewer). We plan to extend the description of the base modelling framework, including the lagged time series approach. Furthermore, we will extend the description of the new introduced datasets such as SEAS5 and EFAS as recommended.
We extended the description of the modelling framework in section 2.1.2 and the new introduced datasets in section 2.1.1 of the track change file.

Regarding the comment Fig 2 and how the different ensembles were aggregated into one CDF, we would like to refer to line 196 ('This was done by computing all the individual daily CRPSS results of all the hindcasts of every station and method, before aggregating the individual CRPSS scores to different temporal scales (weekly in Fig. 2).) and line 155 for the specific calculation procedure ('Equation 1 …was used and the CRPS computed over all ensemble members for each lead day of every hindcast before aggregating it to other temporal scales'). For creating Fig 2 the different CRPSS scores for the different station and method combinations were additionally aggregated by lead day and the average was used for plotting. We added this explanation in line 225 of the track change file.

The reviewer poses a good question regarding what is the advantage of using different ML models, while they only show minor differences in their results. While performing the experiments, we were not expecting to only observe minor differences, while predictions using observed input data did show differences (Hauswirth et al., 2021). The most likely explanation is that the forecasting skill is very much dependent on the skill by which the input variables are forecasted, which apparently make the differences in skill between the ML models insignificant in comparison. We will add an explanation for this in the manuscript. The additional comment can be found in line 239 and section 4 of the track change file.

To keep the manuscript easy to follow we decided to focus on one method in more detail. However, we will add a section on this as this is an interesting aspect to compare to the previous study (Hauswirth et al., 2021), where there were differences found between the models.
The additional comment can be found in line 239 and section 4 of the track change file.

Minor comments:

The acronym LSTM is not defined?
Response: We will define the acronym LSTM (Long Short Term Memory Model) to prevent confusion. Line 153 of the track change file.

Figs.2-6 did not come off well in my black and white copy, but ok for online viewing (which in fairness is probably the most common by now).
Response: We will revisit the figures 2 - 6 and see if we can improve the readability for printed version.

Line 253: Why the 20th percentile? Or why only the 20th percentile? I could imagine that the ability to assess low-flow across a range of severities would be of interest?
Response: We chose the 20th percentile as it is a commonly chosen threshold for low flows and was also interesting to our collaboration partners. Line 193 and 285 of the track change file.

Line 253: Do any of your models include the effect of human interventions and their potential impacts on low flow? For example, water restrictions, operation of control structures to manage low flows etc? If not, is this likely to important in a highly regulated system such as the Netherland's water ways? I think there is so mentioning of this in lines 220-225, but seems to me this is particularly important during low flow?
Response: The Netherlands is indeed a country with high water management infrastructures and plans. We therefore are working with the National Water Authority to include expert knowledge where possible and available. In our previous study we did include a separate water management scenario where we included operational plans of the major water infrastructures. We did see that there was an improvement in our simulations, albeit very minor in terms of our modelling results (Hauswirth et al., 2021).

Line 293: I am not sure I understand how you incorporated water management into your models? Was this done in this study, or is that something that was part of the Hauswirth et al. (2021) study? I think perhaps more detail on this could be included in this manuscript as this seems interesting and important (even if you did not find a strong effect).
Response: In line with the previous comment about water management above: We did a run with water management influence using the same approach as in Hauswirth et al. (2021). This run includes operational rules of main infrastructures which are related to the Rhine discharge at Lobith (one of our main input variables) for two specific input locations and two additional observation records of locations based at smaller infrastructures. We were therefore able to use the same approach regarding the operational rules for the main infrastructures, as these are based on the Rhine discharge we obtain from the EFAS dataset. For the two other additional timeseries climatology was used as operational plans were not available.
We acknowledge that the information given in this manuscript related to water management aspects are very limited and heavily based on the previous study. We will extend the explanation so that the reader will be able to follow better. We added a paragraph in section 3.3 of the track change file.