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). Furthermore, we would like to thank the reviewers for taking the time to read our manuscript.

We would like to specifically thank Louise Arnal for her thoughtful and detailed feedback and suggestions, which were helpful in improving the manuscript. We have carefully considered and responded to each comment and made changes to the manuscript accordingly. In the rebuttal file follows a point-by-point treatment of Louise Arnal’s comments. The response to the comments are given in italic.

We responded on her main points and added a paragraph on the performance of the hybrid framework compared to the EFAS seasonal (re)forecast to the results section.
Furthermore, we added additional information, clarifications and modifications throughout the manuscript based on the minor comments and suggestions to improve the readability.

We realised while doing the rebuttal that the comments were referring to the original manuscript and not the revised one. However, we took the feedback from all rounds into consideration and the submitted track change file is only highlighting the modifications based on Louise Arnal’s comments (previous track change file covers the first round of reviews).

Thank you again for your consideration of our revised manuscript.

Sincerely,

Sandra Hauswirth, on behalf of the coauthors

Review Report #2 – Louise Arnal, louise.arnal@usask.ca

This manuscript presents a hybrid framework for ensemble seasonal hydrological forecasting in The Netherlands, forced with seasonal hydro-meteorological hindcasts from EFAS Seasonal and SEAS5. It shows the performance of the framework for discharge and surface water level hindcasts at several stations and discusses advantages of using a hybrid framework for hydrological forecasting, compared to large-domain dynamical systems. This manuscript is overall well thought-through and presents interesting original findings. However, a few key elements are missing from the analysis to address the objective of the paper more fully. Below are some comments which will hopefully help guide the manuscript revision.

Response: We thank the Louise Arnal for taking the time to review our manuscript and the thoughtful and detailed suggestions that help to improve the research presented here. While working through the comments, we observed that the suggestions and feedback were targeted at the original manuscript and not the updated version after the first round of reviews. Therefore, for this round of improvements we looked at both versions to be sure to include all the feedback from both rounds of reviews. The main comments from Louise Arnal are addressed in more detail below, while a brief response is given for the specific comments in the last section.

Main comments:

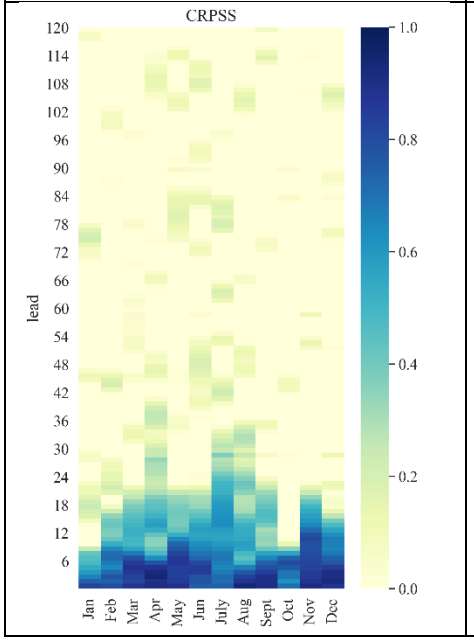
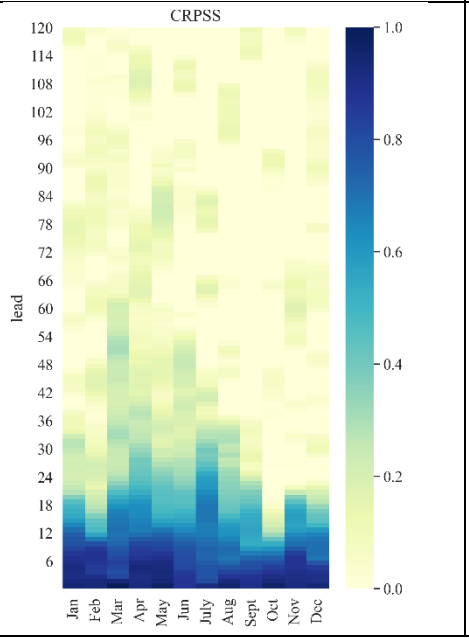
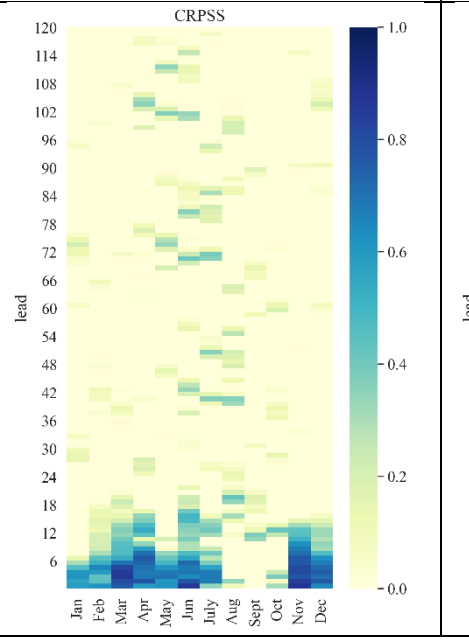
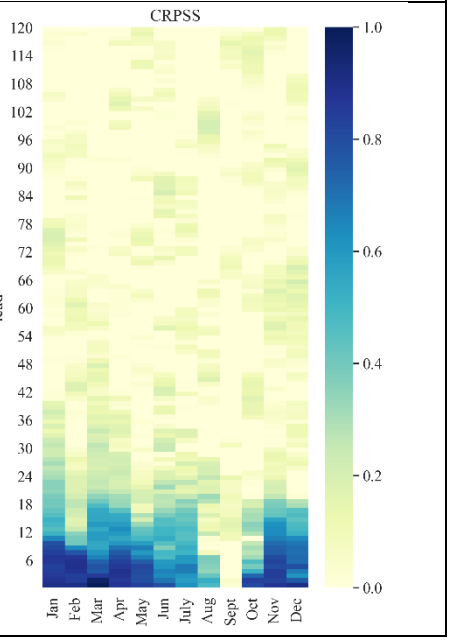
For the first and second comment, we looked at seasonal (re)forecast data, hindcasts and observations and combined it in the CRPSS analysis for two stations. The CRPSS is based on the forecasts and climatology (station observations) and is used to describe the performance of the forecasts/hindcasts. The results are used to answer both comments.

- One of the objectives of this work is to explore the suitability of a hybrid framework for ensemble seasonal hydrological forecasting in The Netherlands, compared to more traditional large-domain dynamical systems. As mentioned in the Discussion, “hybrid forecasting systems have the ability to provide more local information compared to large-scale physically based systems”. To truly show this, it would be interesting to compare the performance of your system to the information already contained in the EFAS Seasonal hindcasts. To do this, two ideas are:
o Using EFAS Seasonal hindcasts as the reference instead of the observed climatology for skill score calculations. The EFAS hindcast values at grid cells that overlap the stations could be used or regional averages of the EFAS hindcast values could be calculated.
o Evaluating the performance of stations downstream of the input stations (Lobith and Eijsden) compared to the performance at those input stations.
- The quality of the EFAS and SEAS5 hindcasts compared to the observations used to train the models is a key element discussed throughout the manuscript. This warrants showing results of the analysis of the pre-processed inputs compared to the station observations.

Response:

To highlight the difference between the EFAS and our hybrid framework we computed the CRPSS for both station Lobith and Eijsden, which are represented by different grid cell in the EFAS dataset. On the following page one can see the improved hindcast skill that was obtained by the hybrid framework compared to the EFAS seasonal (re)forecast. Differences in skill for larger lead times but also throughout different initialization months can be seen for both stations. Furthermore, despite the EFAS data being bias-corrected for the specific locations, the skill for example station Eijsden is relative low for some months indicating that forecasts based on climatology showing similar skill. We added this additional analysis to the results regarding the general performance.

Table 1 Overview of daily CRPSS for locations Lobith and Eijsden for both EFAS input data and hybrid framework hindcasts. An improved skil can be seen for both hindcasts computed by the hybrid framework.

***EFAS** hindcast bias corrected (Lobith)*	*seasonal ensemble **hybrid framework** (this study, Lobith)*	***EFAS** hindcast bias corrected (Eijsden)*	*seasonal ensemble **hybrid framework** (this study, Eijsden)*
			

- Here, you apply different ML models, which appear to give similar results for discharge hindcasts and further results are shown for a single model only. Would you recommend that water managers use a single model or would an ensemble of models be more appropriate. If the latter, could a super-ensemble be constructed from all or a sub-selection of the models? A discussion on this point would be interesting to read in the manuscript.

Response:
We extended our discussion on this aspect previously after the first round of reviews (326-330). Initially, we continued to apply the different ML models from the previous study to see how these models perform in a hindcast setting and whether we would observe similar differences as in the previous study. However, here we found that introducing the uncertainty incorporated by the seasonal (re)forecast information limits the differences we observe between the models. We think though that combining the model results to a super-ensemble could be interesting, especially if the models would for example be more refined for water management where then the potential of all models could be used and incorporated in an ensemble setting. In the discussion we also highlight that if computational demand would be a key factor of choosing models, one might favour some models over the others.

- Did you 1) calculate weekly and monthly averages from the variables prior to calculating verification scores or did you 2) average the daily score values to weekly and monthly timescales? P6 L144-145 and the "pre-processing" step on Fig. 1 suggests approach 1) was used, but P6 L155-158 suggests approach 2) was used. Please clarify in the manuscript. If approach 2) was used, this suggests that you expect to capture the daily variability with a seasonal system. I would expect to see a lot of noise and the final scores to be an average of that noise - but this is just a feeling. If you indeed used approach 2) and changing the method is not an option, could you maybe explore how this affects one set of results for an example station and show results in the Appendix?

Response: We have rephrased and clarified this paragraph in the manuscript and also added additional clarification in the skill score subsections. We first computed weekly and monthly averages prior to calculating scores such as ACC, BS and BSS.

Specific comments:

- Title: Consider adding the terms "seasonal" and perhaps also "ensemble" in front of "hydrological forecasting".

Response: We added these terms to make the title more complete

- Abstract:

o P1 L2: Please introduce the term "flexible" and what it means here.

Response: Introduced the term flexible as it is used and meant in this scope of study

o P1 L4: Clarify that the seasonal (re)forecasting information is obtained from dynamical models to highlight the hybrid nature of this framework from the start.

Response: Added the term dynamical models to highlight the hybrid nature already in the abstract

o P1 L6-7: Please spell out/introduce the acronyms ML, EFAS and SEAS5 here.

Response: added introductions to acronyms

- Introduction:

o P2 L30-31: Instead of “open source” do you mean “openly available data” from these systems? Open source suggests that the source codes used to run these systems are available. Same comment on P4 L106.

Response: We thank Louise Arnal for that clarification and we corrected the terms ‘open source’ to ‘openly available data’ for all instances in line 30, 31 and 106

o P2 L37-40: I can see why you wanted to mention the ESP given the relevance for seasonal hydrological forecasting, but I would suggest removing this sentence completely as it doesn’t appear relevant to your paper specifically.

Response: We removed the sentence as suggested by the reviewer as indeed that information was not relevant to our specific study

o P2 L46: Please elaborate on how a hybrid system offers more flexibility and is less intensive regarding data use. In doing so, I would suggest moving some content from P3 L63-68 closer to the hypothesis, as it provides some useful context that would be valuable to have earlier on.

Response: We added some further additions and moved some content from P3 L63-68 forward for support

o P3 L69: I would suggest removing “data-driven” as these steps are general to all forecasting systems and not specific to a data-driven system only.

Response: We removed data-driven from that sentence to make it more generic to all forecasting systems

o P3 L71: Perhaps add “in near real-time” at the end of the sentence to clarify the distinction with step one?

Response: We added “in near real-time” at the end of the sentence

o P3 L72: Here and throughout the manuscript you use the term “data-driven forecasting framework” to refer to the framework you developed. Would the term “hybrid framework” be more appropriate, as is used in the title?

Response: We agree and changed the term ‘data-driven forecasting framework’ to ‘hybrid framework’ to make it more consistent throughout the manuscript

o P3 L74-76: Please consider removing this sentence as it is not essential here and a bit distracting to the flow of the paragraph, and it is already mentioned in the Methods.

Response: We removed this sentence to avoid repetition

o P3 L76-77: I would suggest rephrasing to “based on their historical performance to forecast discharge and …”.

Response: We adapted the sentence as suggested

o P3 L82: Please consider changing “predictions” to “conditions”.

Response: Replaced ‘predictions’ with ‘conditions’ as suggested

- Material and Methods:

o P4 L99: Could you please mention how many stations were used here?

Response: We added the number of stations for both discharge and surface water level, which were already listed in section 2.1.2 in the updated version of the manuscript after the first round of reviews

For the following points we want to highlight that most of these were already raised in the previous round of reviews. The material and methods section has been updated and restructured and the responses to the comments below are referring to where the information can be found in the updated version of the manuscript after the first round of reviews. Where necessary, more recommendations and suggestions were included

o P4 L104-105: Please briefly discuss in the manuscript why these specific variables were chosen and not others, such as soil moisture or groundwater.

Response: We included a short description on the choices regarding input variables which can be found on line 108 of the update manuscript

o P4 L109: How was the station data extracted from this gridded product? Did you do any downscaling? Same question for SEAS5. Did you extract specific grid cells or calculate spatial averages?

Response: We did select specific grid cells that included the location of the stations Lobith and Eijsden for the input data set. No downscaling was performed, however a bias-correction based on the historical observational records of these stations. We added a small clarification to the text (line 115 and 119)

o P5 L111: Please mention here what input data is used by this module.

Response: The information on the input data is described in the following sentences, where the whole approach is described. Input data to predict the sea level information included u and v wind speed as well as sea level anomalies.

o P5 L114-117: I would find it useful to have a bit more information in the manuscript about the data-processing as I don't understand this step very well.

Response: The information about the data-processing of the seasonal (re)forecasts prior to use in the hybrid framework is included in the following paragraph (line 126-131), which was included after the first round of reviews. We additionally added the comment on grid cell selection as mentioned in the previous comment above.

o P5 L121-142: It would be great to have a bit more information about the model setup. How were the input data used? Was it a mix of data prior to, on the initialization date and over the forecasting period? And with what temporal and spatial resolutions?

Response: The information regarding the model setup can be found in the following section (2.1.2 data-driven model setup, updated manuscript), which gives a summary of the original model setup from a previous study. The handling and pre-processing of the input data, as well as the final input ensemble set, was previously discussed in the data section (2.1.1, updated manuscript), while the model setup (incl. model choices, training and testing, etc.) as well as the current setup for the hybrid hindcasting system is given in 2.1.2. While the original model setup was trained and tested on

historical observations, the hybrid hindcasting system is only run with seasonal (re)forecasting information (without any changes to the model setup)

o P5 L124-125: Please briefly explain how water management aspects are incorporated here.

Response: The incorporation of water management in the original setup was previously highlighted in section 2.1.2 in the updated manuscript after the first review round. Details on how the water management data was prepared for the hybrid hindcasting system was additionally added in section 2.1.1. In principal, operational plans for major water infrastructures gathered from the National Water Authority were used to create additional input data sets (including additional discharge variables), that were incorporated in the original model setup.

o P5 L125-126: How many years of observations were the models trained on?

Response: The models were trained and tested on observational records from 1980-2018, originally (more details can also be found in Hauswirth et al. 2021). In this manuscript the input data was replaced by seasonal (re)forecasting information for the period 1993-2018 (section 2.1.1), matching EFAS and SEAS5 available time periods.

o P5 L127: Add the station names here. I am assuming that these are the Lobith and Eijsden stations mentioned on P9 L234?

Response:
P5 L127: We added the station names regarding the input variables. Furthermore, we added an overview map of the input data after the first review round which can be found in the appendix (Fig A1).

o P5 L127-129: Could you please briefly explain in this manuscript why you tested multiple ML methods instead of a single one? This links back to my 3rd main comment.

Response: We added a brief explanation on why we chose to test all the methods in line 158 and also commented further on it in the 3rd main comment

o P5 L140-141: Please add the number of ensemble members of the final outputs here. At what lead times were the hindcasts produced?

Response: The final target ensemble simulations created by the hybrid framework consist of the same amount of ensemble members (25, 50 after 2017) and lead time (215 days) as the seasonal (re)forecasting input data. The information was added to the manuscript and can be found on line 169.

o P6 L144-145: Mention daily time scales too. I think “bi-weekly” should instead be “weekly”.

Response: we clarified that the evaluation done using skill scores was done for daily, weekly and monthly scale – with a focus on weekly and monthly in the result section

o P6 L158: I suggest changing “baseline” to “reference” in the text to match the equation terminology or modifying the equation to use the word “baseline”.

Response: Adapted the equation corrected ‘reference’ to ‘baseline’ for more consistency

o P6 L161: Please mention the range and best value for the CRPSS. Same comment for the other scores.

Response: We included a line on the range and best value for CRPSS at the end of section 2.2.1

o P6 L169: It is not clear to me from the Methods that the threshold chosen was indeed the 20th percentile and how it was calculated. Is it calculated from the observations?

Response: For the BS the 20th percentile was chosen as a threshold for droughts. This threshold was applied on both observations and simulations.

o P7 L174-175: Why wasn't the climatology used as a reference like for the CRPSS?

Response: We realised that this description was based on one of the earliest versions of our manuscript and does not apply to the BSS calculation that was used for the results described in this manuscript. We incorporated climatology, as we did for the CRPSS. We corrected the description for the latest version.

o P7 L176-184: Please consider providing the ACC formula.

Response: We added the equation for ACC in section 2.2.3

o P7 L178-180: This is a bit hard to follow. Could you please clarify by giving an example?

Response: We removed these lines and replaced them by including the ACC formula to avoid confusion.

o P7 L180: By "compared to the normal correlation" do you mean "relative to the climate values"?

Response: With normal correlation we mean using the standard correlation calculation that is commonly used

o P7 L184: Could you please also mention how many sample points were available for the scores' calculations, especially for the BS.

Response: For the BS we only included stations that had less than 10 missing observation months to ensure that we have long enough records for calculating the skill scores (same selection was used for other scores)

- Results:

o P7 L186-187: Please mention which target variable here.

Response: Included target variable information

o P7 L195-196: Mention that this is for discharge hindcasts.

Response: We clarified that the results discussed here are for discharge hindcasts

o P8 L200: Is the convergence of skill a feature of the ML models? Just curious.

Response: This is an interesting questions. We did not investigate this further but hypothesize that this could also be found for other models, where the forecasting skill decreases with lead time. However, it would also depend on the reference used for the CRPSS and how skilful the reference forecasts are

o Fig. 3: Could you maybe add the main rivers (Rhine and Meuse, which you often mention in the text) to the map for readers not very familiar with The Netherlands. The plots to the right of the map are too small, please make them larger and consider adding the station names at the top of each column. You could also highlight the station Hagenstein Boven on the map using a different symbol or by providing the station code in the text.

Response: We adapted the figure after the first round of reviews and changed the map to a version were major rivers are better highlighted. We additionally included blue lines for the representation of the main rivers. We chose to keep the colour coding for stations instead of adding names as only one station (orange one in this case) will be discussed further on in more detail.

o P9 L226-228: It would be interesting to comment on the noticeable differences between the initialization months from Fig. 4 too, before talking about the differences in more detail on Fig. 5.

Response: Most of the discussion of Fig 4 can be found in the lines above. We added a small addition at the line suggested, however we wanted to focus the discussion on one specific station (also represented in Fig 4) to avoid redundancy

o P9 L236: What is the performance of the seasonal meteorological hindcasts for this time of year? You could perhaps refer to results from the literature looking at the SEAS5 precipitation and evaporation hindcasts quality over The Netherlands/Europe.

Response: Regarding the meteorological forecasting skill, the figure below for temperature was found. One can that the anomaly correlation skill for summer months is around 0 if focusing on the Netherlands, for late summer/autumn months it increases to 0.2-0.4.

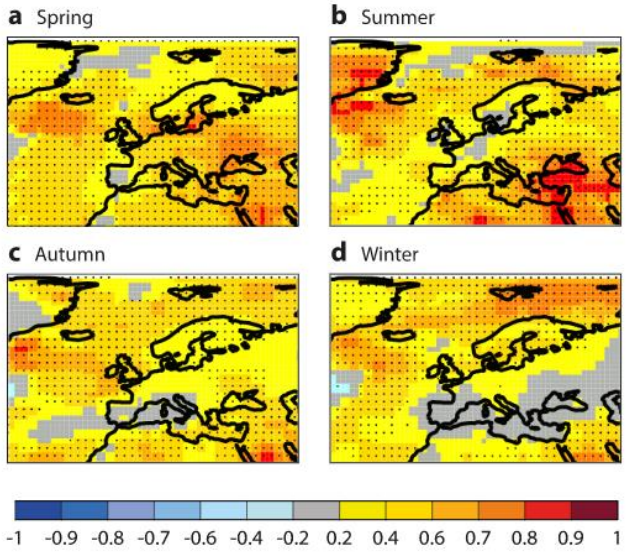


Figure 4 SEAS5 anomaly correlation skill for 2 m temperature in the European region, based on 1981–2016 re-forecasts, for (a) March–April–May, (b) June–July–August, (c) September– October-November, and (d) December–January–February, predicted from 1 February, May, August and November, respectively.

Figure 1 SEAS5 anomaly correlation skill for temperature taken from https://www.ecmwf.int/en/newsletter/154/meteorology/ecmwfs-new-long-range-forecasting-system-seas5

o Fig. 4: Consider making the rivers blue on the maps to distinguish them from other lines. The main rivers Rhine and Meuse could have thicker lines to differentiate them further. You could also display all larger river stations on one side of the map and the smaller river stations on the other side of the map to make the distinction clearer visually and the results more easily interpretable. Why does the colorbar start at 0.4 instead of 0? Same comment for other ACC plots in the manuscript.

Response: We adapted the figure so that the rivers are indicated by blue lines compared to borders in black. Only major rivers are shown. We did not alter the order of the plotted stations as we wanted

to keep the order of the connecting lines to be short and not disturb the map further. For visualisation purpose and ease of reading we only show significant values (indicated by range of 0.4 and higher).

o Fig. 5: Add “initialization month” as an x-label or in the caption, and mention that this shows the ACC for discharge hindcasts.

Response: We included the clarification for discharge hindcasts and initialization month in the caption

o Fig. 6: Same comments as above. I would also suggest removing the leftmost ACC heatmap as it is already shown above. Why is the ACC the only heatmap with empty values? Were all results significant for the CRPSS and BSS? P7 L184 suggests that the same significance criteria were applied to all scores.

Response: We included the ACC map again to highlight all evaluation scores at once and for the same temporal resolution, so that the reader would not have to switch between figures when interested in an overall view of the evaluation scores. For the ACC map only significant values are shown (indicated by range of 0.4 and higher), which means that ACC values lower than 0.4 are not shown/white. The significant values were determined by calculating the confidence level. For CRPSS and BSS the same station selection was considered as for ACC.

o P12 L249: The rank histogram is more of a measure of the hindcast spread compared to that of the observations. The CRPS is a combined measure of the spread of the hindcast and its accuracy relative to the observed value. Please consider rephrasing.

Response: Rephrased the beginning of the paragraph

o P12 L249-250: Is this shown on any specific figure?

Response: This specific comment on the observed pattern throughout several station was not separately highlighted with a figure but found while going through the results of the different stations.

o P12 L257: Except for hindcasts initialized in January.

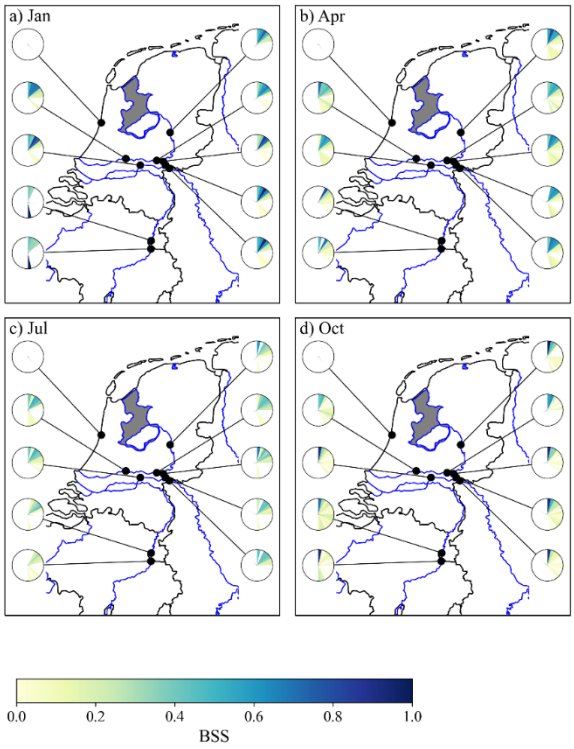
Response: That observation is correct. In the updated manuscript after the first review round, this paragraph has been updated. The BSS for January indicates high performance throughout the first few weeks. A brief explanation of the differences in BSS for various initialization months is given later in line 300-304

o P12 L258-261: Did you consider defining season-specific thresholds instead?

Response: We did not consider this originally.

o P12 L261-262: It would be interesting to see a map of the BSS as well.

Response: Below a map for BSS values, same style as for ACC in Fig 4. Same patterns as in Fig 6 for the example station Hagenstein Boven can be found though stations along the main river network.



o P12 L263: Please briefly mention Fig. A2 here to discuss results from different models, as the section title suggests.

Response: We included a reference to Fig A2 in this section

- Discussion:

o P13 L298: Consider rephrasing to "For hindcasts initialized in the early spring …".

Response: We rephrased the sentences a suggested

o P13 L304: It would be interesting to get an indication of how fast these ML models are for generating such hindcasts.

Response: We added a line to indicate how fast the hindcasts were computed. While the initial training of ML models can take up some more time (depending on the methods and setup used), rerunning the pretrained models can range from a few seconds to minutes for an ensemble member

o P14 L316: This framework still relies on outputs from dynamical systems, which will have to be run (albeit by a different entity/forecasting centre) beforehand. I think it would be worth being more specific and clarifying that your framework offers a computationally frugal way to provide more localized forecast information, which a large-domain system could only do at such a high resolution with a higher computational demand. But that it doesn't cut the computational costs of running EFAS Seasonal and SEAS5 in the first place.

Response: This is a valid and important point and was included in the discussion.