Authors' response to interactive comment by Reviewer #3 Shawn Harrigan

Black text: Reviewer comment

Blue text: Authors' response

Girons Lopez et al. evaluate the forecast skill of theEnsemble Streamflow Prediction (ESP) method using the SMHI operational configuration of the S-HYPE hydrological model for providing seasonal streamflow forecasts across Sweden. They generate a set of ESP reforecasts at 25 ensemble members each for 39,493 catchments, initialised 4 times per month over the 36-year period from 1981-2016. The hydrological model, and reforecasts, are run at a daily time-step and are primarily aggregated to weekly averaged streamflow, out to a 6-month forecast horizon; a number of different temporal averages are also tested, from weekly to 24 weeks. For the 539 catchments with streamflow and water lever observations, an additional simple autoregressive algorithm was applied to correct raw modelled streamflow output prior to generation of the reforecasts. The probabilistic skill of ESP reforecasts was benchmarked using the Continuous Ranked Probability Skill Score (CRPSS) against a probabilistic (25 ensemble member) streamflow climatology (called "historical streamflow"in the paper) with modelled streamflow simulations as proxy observations (called "modelled reality" in the paper), or where available (i.e. 539 stations) in situ streamflow observations (called "observed reality" in the paper), as reference. Results show that ESP is skilful up to 3 months ahead for the most of Sweden. The strength of skill varies widely across the country in space and time and has shown to be linked to a number of key hydrological signatures (15 explored in total). Similar to previous work, ESP skill was highest in slowly responding baseflow-dominated (or high BFI) catchments and least skilful for flashy catchments. Seven unique clusters of similar hydrological behaviour were identified using k-means clustering and ESP skill summarised for catchments within each cluster.

I found this paper very interesting with a comprehensive ESP reforecast experimental design (i.e. long reforecast period, many forecast start dates, large sample of catchments, and cross validation used). It has the clear purpose of providing the scientific foundation for when and where the ESP forecast method is, and importantly is not, appropriate to use in operational seasonal forecasting across Sweden. The paper goes on to explore the potential sources of ESP skill based on correlation with hydrological signatures, while it's arguably a stretch to call correlation a formal attribution, the analysis nonetheless reveals interesting drivers and patterns of skill across the country, including the poor skill from catchments with high human disturbance (e.g. reservoirs). The paper is well structured and written with very good Figures. The analysis presented in Figure 7 is particularly insightful, and an innovative way of presenting forecast skill by clusters of similarly responding catchments. I offer below suggestions on areas where the paper could be expanded and highlight where clarifications are necessary, but these are all minor.

Therefore, I strongly recommend Girons Lopez et al. to be published in HESS. It adds to the growing literature benchmarking the skill of the ESP method with a clear application within operational seasonal forecasting at the national scale in Sweden.

We thank Dr Shawn Harrigan for his valuable comments and suggestions that will undoubtedly help us improve our manuscript. Below we reply to each of these and explain how we will incorporate them into the manuscript.

Main comments

1.) It would be useful to have these parts of the methods expanded/clarified:

a. Pg 3 L89-90: While there is a link to the general website to download the streamflow observations in the "Data availability" section, there is little detail for the reader on if all 539 stations are available in near-real time, which would be necessary to understand the transferability of forecast skill results to operational forecasts in future. Also, are all stations available for the full 1981 to 2016 period for calculation of KGE in Fig. 1 and for calculating the historical streamflow benchmark forecast? Were most of all these stations used for calibrating the configuration of S-HYPE used in the study?

The reviewer raises a valid point here, as hydrological observations are seldom available for long, overlapping periods across a large number of stations. The stations we used in this study are the ones being used operationally (and therefore collecting and sending data in real or near-real time) by the SMHI's service at the time of performing the analysis.

New stations are regularly added and some of the existing ones may be dismantled, but the sample is fairly similar to the one used for model calibration, as a recalibration effort for the most recent version of the model (the one we are using in this analysis) was performed recently.

Station availability and data quality is actually quite complex as, even if a large percentage of stations belong to SMHI, a significant part are external stations. Nevertheless, SMHI performs quality controls on all observations.

In short, because of the reasons listed above, data availability varies greatly among different stations. Nevertheless, most of them have over 20 years of data. Nonetheless, as the reviewer points out, the transferability of forecast skill results to operational forecasts in the future may need to be carefully assessed if there are significant changes in station availability.

In the revised version of the manuscript we will include a figure showing the periods with available observations for all stations (in the appendix) and we will include a sentence in Section 2.1 stating that data are assimilated from different stations at different times since data availability is not the same throughout the different stations for the 1981-2016 period.

b. Pg 5 L116-123: I find the AR correction interesting, but there is very little detail on how it was applied within the current experimental design, and perhaps even if it was implemented in such a way that is as consistent as feasibly possible to the configuration that is/will be implemented operationally?

The implementation of the AR correction in this reanalysis is indeed as close as possible to the operational setup. For instance, let us consider the case of a catchment which has observations throughout the analysis period. For each ESP initialisation, the model outputs are corrected up to the day before forecast initialisation. Then, as observations are theoretically no longer available, the model output correction starts from the latest correction value and exponentially decreases with time until the model outputs become clean simulation results (the rate of decrease is controlled by a model parameter). In the revised version of the manuscript, we will include this paradigmatic detailed description of the AR method.

c. Pg 5 Sect.2.3: The exact reforecast size is mixed between Sect. 2.2 and Sect. 2.3 and the reader has to try piece it together, it would be good if summarised. My understanding is that the reforecast dataset used has the following size: 39,493 catchments; 1728 start dates (4 start dates per month x 12 months x 36-year reforecast period (1981-2016)); weekly averaged streamflow out to 6 month forecast horizon at 25 ensemble members each?

That is correct. In the revised version of the manuscript we will rework the methods section to ensure that the description of how forecasts are generated is contained in a single section (Section 2.2). We will consequently also rename this section to "Hydrological modelling and forecasting".

d. Pg 5 L130-134: I think it could be confusing to refer to the probabilistic streamflow climatology benchmark forecast as "historical streamflow" because historical streamflow could more generally be interpreted by readers as the reference observations. I think it's more informative to be explicit about the type of benchmark forecast used for benchmarking skill (here, you indeed choose climatology which is the most appropriate given the seasonal forecast horizon).

We agree with the reviewer in that terminology should be used in a restrictive sense, as otherwise it could lead to misinterpretations. We will ensure that the appropriate term (streamflow climatology) is used throughout the text in the revised manuscript.

2.) Pg 7 L172-173: I'm not sure this is the correct conclusion from my interpretation of Fig. 2b and Fig. 3. It looks like skill initialised at the start of March (light green in Fig. 2b) is higher than any of the winter months, at least for the 1 week forecast horizon. This is confirmed in the map for 1 March in Fig. 3 for 1 week. Can you please clarify?

That is correct. The highest skill is indeed achieved for reforecasts initialised on 1 March (CRPSS above 0.8). The period with highest skill for forecast week 1 is actually between 8 December and 1 March, which we simplified in the text as "for forecasts initialised in winter" (Line 172). After 1 March the skill already decreases noticeably. This may be explained by the hydrological regimes of a large part of Swedish catchments, which generally start to increase in April-May, in addition to the general lack of precipitation in winter and early March. We will modify these lines to be more accurate.

3.) One of the key advantages of benchmarking ESP over Sweden is the opportunity to explore the role snow accumulation and melting has on controlling ESP skill. I can't help but think there's an additional piece of the puzzle missing in attributing ESP skill. While hydrological signatures are useful, e.g. baseflow index (BFI), there is not much discussion in the paper on the hydrological

processes within those catchments that are the source of ESP skill, based on information content and hence memory in the initial hydrological conditions. For example, a key question missing from the analysis is do catchments with a large contribution of streamflow from snow melt provide high skill when initialised around the snowmelt season? In practice, a catchment can have a high BFI due to several slowly responding processes (e.g. large groundwater/soil storage, snow, lakes, or a combination). I do not request this analysis is done, but it would be good to hear the authors' opinion and perhaps it could be worked into the discussion on the (initial hydrological condition) sources of ESP skill in Sweden.

We thank the reviewer for this interesting point. Unfortunately we did not calculate the contribution of streamflow from snow melt, and hence we cannot explicitly explore the role of snow accumulation/melting on ESP skill. Only in a case study investigation over the Umeälven river basin (snow dominated and heavily regulated system for hydropower production), we recently showed that assimilation of a snow water equivalent satellite-based product, particularly over the winter and spring seasons, significantly increased the streamflow forecasting skill. Note that this is still unpublished, whilst a manuscript is under preparation with an expected submission in early 2021.

Moreover, we agree with the reviewer that the definition of river memory can be a combination of processes, such as groundwater/baseflow contribution, snow accumulation/melting, and hydrograph dampening from lakes. Snow processes tend to define the river memory only seasonally (for example, precipitation in the form of snow in early December will be accumulated and further released as melting during the spring flood period), and hence the role of snow on ESP skill is expected to have a seasonal pattern too. This view will be mentioned in the last paragraph of Section 4.1 (Discussion).

Technical comments

4.) Pg 3 L69: Not sure "spread to other actors" is clear. A suggestion is: "ESP seasonal forecasts are produced operationally but have not been used widely in real-world applications due to lack of information on their skill...", or something to that effect?

We thank the reviewer for his suggestion. We will change this passage in the revised manuscript accordingly.

5.) Pg 4 L96-98: Can you please confirm the time scale the KGE was calculated, I presume it was calculated at daily time step from 1981-2016?

That is correct, the KGE values were calculated based on a forward run at a daily time step for the entire analysis period (1981-2016) using the S-HYPE model without station correction. We will clarify this in the revised version of the manuscript.

6.) Pg 4 Fig.1: Could you please add into the caption or text what exactly is shown in Fig. 1b and c in the coloured shapes, I presume it's the river network, or is it the river network downstream from an observed gauge only?

The coloured shapes in Fig. 1 correspond to those catchments that are being significantly corrected by observations over the entire analysis period (Fig. 1b) and that have a significant degree of

regulation (Fig. 1c). Even if they do not show the river network directly, they correspond quite well with it, since most stations and dams are located along watercourses. One can actually see the difference between both at the border with Finland (north-eastern part of the country): even if the Torne river there is not regulated (it is not shown in Fig. 1c), model outputs are still corrected - yet to a small degree - using the observations gathered from the stations along the river (which can be seen in Fig. 2b). We will clarify this in the text.

7.) Pg 20 L365: Suggest changing “reliable” to “skilful”, as reliability was not explicitly evaluated.

We thank the reviewer for pointing this out. This is correct and we will therefore ensure that the appropriate term (i.e. skilful) is used throughout the revised version of the manuscript.

8.) Pg 21 L396-399: “sys”, “bench” and “pft”more typically subscript, not superscript (i.e. $CRPS^{sys}$ should be $CRPS_{sys}$). Also, CRPSS values can range from 1 to $-\infty$, not “low negative values”.

We will include the proposed modifications to the revised version of the manuscript.