Review of Girons Lopez et al.

Girons Lopez et al. evaluate the forecast skill of the Ensemble 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.

Main comments

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

a. **Pg3 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?

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?

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?

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).

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?

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.

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?

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

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?

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

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".