Response to reviewer 2

Reviewer's comments are in blue, our comments are in black.

Brief Overview

The paper presents a very interesting study related to the implementation of a prototype for seasonal forecasting in Swedish rivers based on hydrological modelling and seasonal meteorological forecasts. The prototype is compared to a traditional operational EPS approach and to climatology. Results show benefits in the use of the prototype. The paper is well written, methods are adequately described, and assessments seems suitable to the objectives. I have only a few major and minor comments about the manuscript.

Major Comments

P2, l20-25: Please observe that it historical observations are referred two times. And only in the second one it is presented as the ESP approach. The explanation here could be better.

We have reworded page 2, lines 19 and line24, to make this clearer. They now read as follows:

"…and then force it with either historical observations (called ensemble streamflow prediction or ESP; e.g. Day, 1985)…"

"Another dynamical approach is the well-established ESP method (Day, 1985)."

Evaluation section: I understand that one of the limitations of the work is that authors were not able to evaluate properly the ensemble, since most of the used metrics are related to transforming the ensemble into the ensemble mean, and then evaluating it as a deterministic forecast. Authors did not even experiment testing some other metrics?

With only 35 data points per station, one data point per year, we felt that it was not enough data on which to perform a robust probabilistic evaluation on. We experimented with the metric CRPS but were ultimately uncomfortable presenting those results due to their uncertainty arising from the limited data used in the analysis. We should also point out that the inter quartile range skill score (IQRSS) and uncertainty sensitivity skill score (USS) used in this work are basic ensemble evaluation metrics and, although not a full probabilistic evaluation, do give some insight into the performance of the forecast ensembles.

P5, l10: It is relevant to better explain what is the data used in the bias correction. Also, I think this procedure has great impact in results, but it is not adequately described. My suggestion is to explore more this point.

We have expanded our description of the bias adjustment and have replaced page 5, line 10 with the following:

“A change to previous work has these daily P and T data bias adjusted first before being used to force HBV. The bias adjustment method used is a version of the distribution based scaling approach (DBS; Yang et al., 2010) which has been adapted for use on seasonal forecast data. DBS is a quantile mapping bias adjustment method where meteorological variables are fitted to appropriate parametric distributions (e.g. Berg et al., 2015; Yang et al., 2010). For precipitation, two discrete gamma distributions are used to adjust the daily seasonal forecast values, one for low-intensity precipitation events (≤ 95th percentile) and another for extreme events (> 95th percentile). For temperature, a Gaussian distribution is used to adjust the daily seasonal forecast values.

Observed (Sect. 2.6 Study area and local data) and seasonal forecast (Sect. 2.7 Driving Data) time-series of P and T spanning the relevant forecast timeframe (e.g. Jan-Jul for forecasts initialised in January) and for the reference period 1981-2010 are used to derive the adjustment factors to transform the seasonal forecast data to match the observed frequency distributions. First the precipitation data is adjusted then the temperature data. The latter is done separately for dry and wet days in an attempt to preserve the dependence between P and T (e.g. Olsson et al. 2010; Yang et al, 2010). Adjustment factors are calculated for each calendar month as the distributions can have different shapes depending on the physical characteristics of the precipitation processes that are dominant. It should be emphasized that the adjustment parameters were estimated using much of the same data to which they were applied. Ideally the parameters would be estimated using data that does not overlap the data which is being adjusted. However, this was not possible in the scope of this work.”

Conclusions:

Authors commented that the prototype was put into operation as a beta product at SMHI in January 2017. This gives openness for another discussion: in an operational perspective, are the benefits verified for the prototype enough to justify the implementation? I understand that yes, but also the prototype is more dependent on data and require more processing power and time to run, right?

Yes, we and the power companies think that they do. It must be emphasised that every percent improvement in the forecast error can potentially be converted into large financial revenues for the power companies and energy traders. So an average improvement in forecast error, over all subbasins and initialisation dates, by 6% (individual results can be as high as 31%, see figure 4) can be viewed as a significant improvement. Care was taken while developing the prototype to minimise the added computational power and data requirements. Additionally, these forecasts are made only once a month so the additional computational time, ca. 1 extra hour, is not a significant factor.

Page 16, line 6 was rewritten to emphasise that the implementation of the prototype as a beta product was done together with the power companies. It now reads:

“These results have been met with great interest from the hydropower industry and the prototype was put into operation, in cooperation with the power companies, as a beta product at SMHI in January 2017.”

Minor Comments

P1, l16: “considered” is doubled in the text

The first instance has been deleted so that it now reads, “Both the considered multi-model methods considered showed skill over the reference forecasts…”

P2, l34: Please explain better what is “limited success”. Only one case is cited

The paper cited is a review of the different experiments performed at SMHI to improve the forecast error of the SFV. They found that a despite these efforts the

P15, l19: The sentence is confusing. Please revise.

The line has been reworded to read, “The IQRSS values show that the prototype tends to produce sharper forecasts than HE early in the season i.e. for forecasts initialised in January and February in cluster S^1 and forecasts initialised in January, February and March in clusters S^2 and S^3. This is reversed for the remaining initialisation dates where HE tends to produce sharper forecasts than the prototype.”

P6, l10: “subbasins sub-basins”

The second hyphenated instance has been deleted.