Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-542-RC1, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden" *by* Marc Girons Lopez et al.

Louise Arnal (Referee)

louise.arnal@usask.ca

Received and published: 25 November 2020

In this paper, the authors evaluate the performance of an ensemble streamflow prediction (ESP) hindcast dataset for seasonal streamflow forecasting in Sweden, produced with the S-HYPE hydrological model driven by resampled historical meteorological forcings. They look at the ESP hindcast skill against a benchmark, historical streamflow climatology, for 39,493 Swedish catchments. They overall found that the ESP is skilful up to 3 months ahead in Sweden, but that the skill varies in space and time, depending on: the aggregation period selected, the catchment's hydro-climatic characteristics and regulation. They analyzed the skill against hydrological signatures, clustering basins in

7 geographical clusters in Sweden, and found that higher skill values are associated with baseflow-driven catchments.

This manuscript is overall well-written and the sound methodology leads to valuable findings both for research and for operational streamflow forecasting in Sweden. Since the focus of this manuscript is on operational forecasting to guide decision-making, further context and discussion around the potential impacts of these findings on operational decision-making is crucial. Below, please find specific comments which I hope will be helpful in shaping this manuscript further for publication.

Specific comments

Section 1:

- P1 L27: "Even if most day-to-day decisions on water-related issues are based on short- and medium-range forecasts, some activities, such as water reservoir operation and optimisation or strategic planning, benefit from long-term forecasts." Do you have any quote or public material you could share about needs of reservoir operators in Sweden? It would help emphasize the user-oriented aspect of your paper.

- P1 L29: "Despite their inherent uncertainties". I wonder if you could very briefly here cite a few examples of the uncertainties you refer to, for readers less familiar with forecasting on longer timescales?

- P2 L34: I think it is important to cite Day 1985 here (Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. J. Water Resour. Plann. Manage., 111, 157–170, doi: https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)).

- P3 L63: "The Swedish Meteorological and Hydrological Institute (SMHI) has long been operationally providing streamflow forecasts and hydrological warnings to relevant actors in hydrological risk management (municipalities, county boards, Swedish Civil Contingencies Agency), as well as to the general public." Please clarify that this is for Sweden.

- P3 L69: “ESP seasonal forecasts are produced but not generally spread to other actors due to uncertainties in their skill and interpretation by external parties.” This is an interesting comment and I wonder what system actors currently use for prediction on such timescales in Sweden? Please consider mentioning this in the introduction to provide some further context.

- P3 L72: “In terms of regionalisation, four main hydro-climatic regions based on hydro-climatic patterns (Lindström and Alexandersson, 2004; Pechlivanidis et al., 2018) have typically been used for water management in Sweden. However, these regions were not put forward with consideration to seasonal streamflow predictability over Sweden and might therefore be of limited use for this purpose.” This appears a bit out of context here, please consider moving to the Methods section instead.

Section 2:

- P3 L86: When you say “measured values from all available stations” do you mean station observations? Please clarify here. Same for discharge and water level data. Please clarify that these are observations.

- P3 L91: Is HYPE distributed, lumped or semi-distributed? And how were the meteorological inputs prepared (e.g. interpolated) for the model to ingest?

- P3 L93: It is unclear to me at this stage how an “analysis of model outputs” was performed for 39,493 catchments if you only have 539 observation stations? Please clarify here.

- P4 L96: Please provide the lowest and highest score possible for the KGE for readers not familiar with this performance metric. Out of curiosity, has a S-HYPE model evaluation been published that you could refer readers to?

- P4 L100: I suggest putting figures 1a-c in the same order as they are mentioned in the text. I was slightly confused and thought I had missed explanations about 1b, which in fact come after 1c.

- P4 L108: “Nevertheless, since dam operation is continuously adapted (within certain bounds) to the present and most probable future meteorological and hydrological conditions, these general regulation regimes are expected to be of little benefit for seasonal forecasting purposes.” This is a big statement which warrants further investigation (not necessarily in this paper though!).

- P5 L111: It may be worth explaining further how the ESP hindcasts are produced – i.e. how initial hydrological states are produced to initialize the model for each forecast start date, each meteorological forcing year corresponds to a streamflow hindcast ensemble member, etc. Perhaps a schematic would help make this clear to readers not familiar with the ESP. I also wonder what the lead time of these hindcasts is?

- P5 L129: “as a station-corrected simulation approach was used to achieve the best possible initial conditions.” I am not sure to understand how a station-corrected simulation approach was used for catchments without station observations? Please clarify.

- P5 L130: Do you know if users in Sweden indeed use “ensemble forecast based on historical streamflow”?

- P6 L151: Could you please provide some more information about the k-means clustering method, or refer the readers to publicly available material further explaining this method?

Section 3.1:

- P7 L156: Please introduce Figure 2 prior to commenting on the results. What do the plots show and what is the highest/lowest score possible for the CRPSS? Same for subsequent figures.

- P7 L156: By lead time, do you mean the aggregation periods mentioned on P5 L142? Or are the results in Figure 2 from daily outputs, and up to what lead time? Please clarify here and in the Figure caption.

- P7 L162: I am not sure to understand what you mean by “the common monthly

initialisation frequency of climate prediction systems". Could you please further explain or reword?

- P7 L163: "By increasing the frequency of forecast initialisation (e.g. from once a month to once a week), and hence frequently updating the initial hydrological states, it is possible to maintain a high streamflow forecast skill for extended forecast horizons". This is a very interesting finding and I wonder if you could comment in the Discussion on how it could be translated into operational decision-making? E.g. Would decision-makers be willing to alter their decisions regularly with each forecast initialization/update?

- P8 L184: I am not sure where these lakes are in Sweden. Perhaps it would be helpful to add a map of Sweden with a few key geographical indicators (e.g. elevation, lakes – with legends for the lakes you refer to –).

- P8 L188: While I can see lower skill for the regulated rivers, it is hard to identify which rivers you refer to on L191-192. Another plot, such as a zoomed in plot, might be necessary to show these results more clearly.

- P8 L191: "future trends in streamflow". This sounds like you are looking at events (e.g. high/low flows). It is perhaps better to rephrase to "future streamflow".

- It is clever to aggregate forecasts for different periods (Figure 3). This enables to retain some skill for longer lead times than otherwise possible when looking at Figure 2. I wonder if users are interested in such time aggregations, or if they would prefer weekly/monthly aggregations instead? Could you perhaps comment on that in the Discussion, as this is important for the user-oriented analysis you are trying to achieve.

Section 3.2:

- Figure 4:

- Before looking at this figure, it wasn't clear to me that the analysis was performed for different aggregation periods as well as lead times. Could you please clarify this in the

Methods section?

- Could you please add ticks (and perhaps tick labels where possible) to all subplots of this figure as it is difficult to follow the results clearly without.

- Do you have an explanation for the sudden increase in skill for hindcasts initialized on 1 March, with a 8- vs 12-week aggregation period? Is it because you are predicting streamflow for the summer with the 12-week aggregation period, which is "easier" to predict as levels are generally low during this season? Please consider reflecting on this briefly in the paper.

- P10 L198: Could you please remind us here which aggregation periods were used for this analysis?

- P10 L203: "Even if, as expected, forecast skill decreases when forecasts are aggregated over long periods, a comparatively higher skill is maintained over longer time horizons than when forecasts are aggregated over short periods." It would be interesting if you could add an indication of the lead time at which the skill is 0 for shorter aggregation periods (results from Figure 2) on this figure.

Section 3.3:

- I would argue that results for longer forecast horizons would be good to show as well as the focus of this paper is on seasonal forecasting. Perhaps correlations could be stronger when calculated against another performance metric which might not weaken so much over time (e.g. CRPS instead of its skill score)?

- To what extent do you think these results are dependent on your hydrological model? Please consider commenting on this in the Discussion section.

- Could you please increase the font size of the correlation coefficient on each subplot of Figure 5? It took me a bit of time to notice them.

Section 3.4:

- Table 2: It would be good to show the range of elevation, annual precipitation, etc. instead of just the mean values, to show the catchments variability within each cluster region. This might become a bit messy and could be clearer in a figure rather than a table.

- P14 L241-254: It may be easier to follow by having these observations as bullet points in Table 2. It might also make it easier to link the results presented in Figure 7 with the cluster characteristics.

- Could the large/small spread in forecast skill shown in Figure 7 be caused by large/small basin differences within these clusters? E.g. spread in the topographic, climatological or hydrological characteristics (from Table 2) within each cluster. It would be interesting to hear your thoughts on this here on in the Discussion. For example, cluster 5 catchments appear more spread out throughout Sweden (Figure 6b) compared to cluster 6 catchments.

Section 4:

- P18 L306: "forecast initialisations are not expected to provide an added value to the forecast service." I would argue the opposite. You have shown in your paper that more frequent forecast initializations could substantially increase the forecast skill. The added value is potentially immense for decision-makers. The challenge remains to translate this into actionable outputs for the users, as you mention it briefly. Please consider rephrasing and elaborating on this.

- P19 L332: Would you be able to add a figure to the paper to support these very interesting findings?

- P19 L344: "Skilful ESP seasonal forecasts for these rivers should allow for early planning and allocation of resources that could greatly contribute to mitigate potentially severe ice break-ups." To evaluate this, a different performance metric, such as the brier or ROC score for high flow events, might be better adapted than the CRPS. Do

you plan to look at this in the future?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-542, 2020.