Review of 'Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments' by Trine Hegdahl et al.

Jan Verkade, November 2018

Overall impression

This manuscript is suitable for publication. The research described in it has a clear objective which is to try and determine if 'calibrated temperature ensemble forecasts' result in better streamflow forecasts compared to the non-calibrated equivalents. The research setting, the approach and the data used is well described and the results are well laid out. I have a few concerns/questions/suggestions but these would require only minor revisions to the manuscript.

Minor comments

Overall

- Multiple references are made to seasons in which the effect of temperature forecast calibration on streamflow was negligent. You're right to point out that the reason is that temperature forecasts only matter if/when it affects the simulation of snowmelt processes. You could consider mentioning this in the start of the paper, explain that for this reason, you're looking at only those seasons where temperature affects streamflow through either rain-falling-as-snow or through snowmelt, and then omit reference to the other seasons altogether. I find it a bit distracting from the main points.
- For many hydrologists, the word 'calibration' has a different meaning from how it is used in your paper. I acknowledge that your meaning is consistent with how many meteorologists would interpret it. I would recommend to address this issue by either use a different word (I believe HESSD readers may be more familiar with 'post-processing') or by addressing this in the text somewhere.
- Citations aren't always properly formatted. I think I've seen ((double parentheses)), for example. In S3.1.2, l12, a correct way to refer to the evidence would be (Seierstad, 2017) with the 'personal communication' listed in the bibliography. I think. I've also seen citation in which both first and family names are listed. May be good to verify against Copernicus citation rules.

Abstract

- l9-11 These sentences distract from the point you're going to make. While the facts you state may have a place in the introduction, I would omit these from the abstract.
- l20 'the HBV model is used to *calculate* streamflow'. The verb *to calculate* presumes certainty. Pls consider using *estimate* instead.
- l21 'influenced'. My understanding is that 'influences' (and the associated verb) are a thing of the mind ("Who are your main influences?" "Joan Baez"). For physical processes, I think 'affected' is more suitable.
- l26 'however'. I don't think this sentence contradicts anything that was stated before. Hence, the word 'however' may be omitted.

Section 3.1.2

- I am not entirely sure who provides the calibration parameters. L5 suggests MetN, but the sentence "To establish the calibration parameters..." (l8) may be interpreted as an explanation of how the authors have done this.

- In the Met Norway procedure, why aren't temperature *observations* used? Are the HIRLAM reanalyses deemed to be sufficiently certain? This may deserve a few informed comments.
- If I am correct in understanding that both the raw and the calibrated ensembles have been provided by Met Norway then maybe this should be stated more clearly. Or is it the case that Met Norway computed the calibration parameters on a data set from 2006-2011 and that you applied these yourself to a data set ranging from March 2013 through Dec 2015? If so, maybe state this more bluntly?
- I am assuming that you used a HIRLAM reanalysis. Is that correct? If not, what lead times are you using and do the HIRLAM forecasts you used have the same max lead time as the ECMWF ensembles? I am only familiar with a few instances of HIRLAM and these all go out to just over 2 days max.
- By off-setting Tens against Tcal, you create the impression that Tcal is not an ensemble forecast. Consider using Traw and Tcal instead.
- l29-30. The 'assessment' was done by you, not by the ensemble range.
- On assessing sharpness: how confident are you that a visual assessment does the job? Pls consider plotting the empirical distribution of sharpness of all your forecasts and comparing those.
- If you're calibrating the temp ensembles on a leadtime by leadtime basis and on a grid cell by grid cell basis, chances are that you'll change the temporal pattern (forecasted temperature as a function of time) as well as the spatial pattern. Does this in any way affect use in streamflow forecasting? I believe there are some techniques that may be helpful in trying to restore spatial-temporal relations (the Schaake shuffle springs to mind). Would these have a use in present study?

Section 3.2

- Would it be fair to say that temperature forecasts are only relevant if they can discriminate between freezing and non-freezing situations? If so, would it be justified to focus more on this discrimination? Perhaps by defining an event (T<0, for example) for which one can compute a range of verification scores (false alarms, hits, ROC, Brier's probability score, etc). I acknowledge that this would be feasible for temperature and less obvious for streamflow.

Section 4

- " To reduce the amount of presented results, the remaining part of this paper focuses on CRPSS for a lead time of 5 days." This is fine, but temperature forecast at 5-day lead time may not affect streamflow forecasts until a (much) longer lead time. Or conversely, streamflow forecasts at day 5 would have been affected by a day 2 temperature forecast (this is an example). As in some cases you're comparing Q-forecasts with T-forecasts, how have you accounted for this?

Section 4.1

- In the text, you refer to observed temp as To. In plots, as Tobs. Pls make this consistent. I recommend using Tobs throughout.
- L23-25. These sentences are better placed in a discussion section, I think.
- L19 'influence' is missing an 's'. Pls consider replacing by 'affects' though.

Section 4.2

"Scatter plots of the difference between CRPSS for calibrated and uncalibrated forecasts". CRPSS in itself is a fairly abstract measure. The difference between two CRPSS scores is, I find, even more abstract. What's the meaning of those values? As CRPSS is a skill of a forecast versus a baseline, why not simply calculate the CRPSS of the calibrated forecasts using the CRPS of the uncalibrated forecasts as a baseline?

Section 5

L7: 'dispersion' is not an expression of quality but a characteristic of an ensemble. Saying 'dispersion improved' makes little sense then?

Section 5.1

- L11 "skill... depends". Consider replacing by "skill... varies with".
- "Quantile mapping is sensitive to forecasts outside the range of calibration values and period". I think it would be good to point out that this is true for any statistical post-processing procedure.
- Immediately following: "and can be a" –> "and *this* can be a"
- On the causes of temperature forecast bias. You go into some detail to explain a situation in which land is colder than sea. Would this be a typical situation for summer/winter? If so, can you more directly link this to some of the results you're showing?

Section 5.2

L10 Grammatically, this sentence is awkward if not wrong.

Figures

Overall

Many figures use a lot of white space between various plots/panels. Consider reducing this or, even better, removing altogether.

Figure 1

- Do the grey polygons add up to 139 in total? If so, many must be **really** small?
- Caption: consider using 'boundaries' instead of 'limits'

Figure 4

- Why plot the ensemble **mean** and not all five ensemble members, possibly as horizontal lines?
- The axes of the plots in the right-hand column vary. Please consider unifying this. Also: please consider ensuring that horizontal and vertical axes are identical. Maybe they are, but the labeling isn't.

Figure 5

- What lead time are these plots for?
- Is the lead time for T identical to that for Q? What is the 'response time' of the catchment to snowmelt? If not zero then shouldn't this be taken into account somehow?

Please consider...

- ... removing data for seasons for which temperature has little or no effect on streamflow levels.
- ... unifying horizontal and vertical axes. it took me a little while longer than I cared to realise that the light grey slanted line is the 1:1 diagonal.

Figure 6

- What do you want the reader to compare? CRPSS(T) and CRPSS(Q)? Or CRPSS(spring) v CRPSS(autumn)? Pls ensure panels are ordered accordingly.
- pls ensure that within a row, panels have identical vertical axes so this comparison can indeed be done (i.e. the reader can then easily compare the top left with the top right plot)

Figure 10

- The background colours have an effect on the colouring of Qens and Qcal. Please consider removing the background shades. Maybe replace these by threshold lines only?
- Please consider removing the number of lines in the plot, for example by only showing a shaded area with no line at the edges thereof.
- What is the purpose of showing both the 'real' observations and the 'model streamflow with SeNorge observations'? Is this distinction made in the paper, and addressed?
- Consider reversing the order of the graphs. The 9d lead time graph was available before the 2d lead time graph?
- The horizontal axis labeling is not in English.
- As all horizontal axes are identical, pls consider removing white space between plots altogether and only label the axis of the bottom plot.
- The warning levels aren't relevant, are they? On reflection: you're scoring the forecast ensembles using CRPSS and rank histograms. This shows absence of preference for doing well for 'extremes', even though the work appears to be inspired by forecasting for floods. How is this consistent? Maybe omit references to 'floods' altogether?