Thank you for the positive and good evaluation of our article. We appreciate the comments that are valuable and helpful in order to improve the manuscript.

We would like to apology for the missing references. The error emerged when we specified the HESS format, and un-intentionally deleted many references from the reference list. The main author should nonetheless have detected this flaw prior to posting.

Replies and corrections are done as follows: the Author responses (AR) are marked with red text, while the author's suggestions to corrections (AC) are marked with blue text. All Referee comments are kept black; we use page and line number when needed to specify the appropriate location.

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.

AR: This is a good suggestion. We think, however, that it is useful to include all seasons in the first part of our analyses in order to highlight the differences between seasons, which subsequently provide the motivation for leaving some seasons out of the final analysis.

AC: No changes will be introduced in the manuscript.

• 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.

AR: We agree that hydrologists might interpret the term "calibration" to "hydrological model calibration", and we will clarify our use of the terminology as illustrated in Figure 2. Pre-processing is, in our paper, a general term for any modifications applied to a raw meteorological forecast. We distinguish between calibration and downscaling, that both are pre-processing methods. This is consistent with the terminology used by the Norwegian Meteorological Institute (MetNorway) (https://github/metno/gridpp).

AC: We will clarify the distinct use of the term calibration in this paper.

• 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.
AR: Thank you.
AC: The citations and references will be formatted according to the HESS standard.

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.
AR: You are right. We will consider rewriting the abstract.
AC: We will change the first sentences as follows:
"In this study, we used meteorological ensemble forecasts with the hydrological models to quantify the uncertainty in forecasted streamflow, with a particular focus on the impact of ensemble temperature forecasts. In catchments with seasonal snow cover, snowmelt is an important flood generating process."

• l20 'the HBV model is used to *calculate* streamflow'. The verb *to calculate* presumes certainty. Pls consider using *estimate* instead.
AR: Thank you, we will change as suggested, i.e. using 'estimate' both in the abstract and in the text.

• 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.
AR: Thank you. We will change 'influence' used as a verb to affect, and to 'effect' where 'influence' is used as a noun.
AC: Change to affects or in some cases effect: p1 l23; p2 l5;p8 l26 (effect); p9 l29; p12 l25(effect); p13 l25 (effect); p14 l7 (effect)

• l26 'however'. I don't think this sentence contradicts anything that was stated before. Hence, the word 'however' may be omitted.
AR: Thank you, we will omit "however".
AC: "Altogether, it is evident that temperature forecasts are important for streamflow forecasts in climates with seasonal snow cover."

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.
AR: MetNorway did the quantile mapping, and established the calibration parameters. The calibration parameters were originally used to bias correct the temperature forecasts as provided on yr.no (the Norwegian weather forecasting). We applied the Met-parameters to the raw ENS temperature forecasts of our selected period.
AC: We will change the sentence in l8: "To establish the calibration parameters MetNorway used both ENS re-forecast (Owens, 2018) and Hirlam data from July 2006 to December 2011 interpolated to a 5×5 km^2 grid."

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.

Author Response to RC#3 Jan Verkade

AR: You are right to point out these differences in data sets used for calibration of forecasts and the hydrological model. First, as you mention, SeNorge and Hirlam are not the same data. Hirlam is a short-range regional forecast model (4 km resolution) used in the operational weather forecast for the first 2 days, whereas SeNorge is a dataset where observations are interpolated to a 1 km grid.

In this study, we wanted to use the available operational method from MetNorway, and they use quantile mapping with Hirlam as a reference to calibrate the ECMWF ensemble forecast. Both Hirlam (for the first 2-3 days) and ECMWF (for the following 7-8 days) forecasts are used in the operational weather forecast (yr.no). Using Hirlam data to calibrate ECMWF will improve the transition between the forecasts. Hirlam is available as a sub daily grid and makes it possible for MetNorway to provide different calibration parameters for day and night, whereas SeNorge is only available as a daily grid and would not offer this possibility.

Hirlam have less errors than ECMWF in the temperature forecast for Norway (Engdahl et al. 2015), and as we see from e.g. fig 6 and 7 that the calibration improves especially the cold biases in the ECMWF forecasts. When we evaluated the hydrological model, the temperature calibration improved, in most cases, the hydrological forecasts, providing an indirect confirmation that the HIRLAM temperature is less biased than the ECMWF temperature. Nevertheless, the results suggest that there might be improvements using the SeNorge data instead of Hirlam, but this needs to be tested (beyond the scope of this study).

AC: We will rewrite the following sentence: "MET Norway uses Hirlam (Bengtsson et al., 2017) temperature forecast (on a 4×4 km^2) to provide a reference for the parameter estimation (calibration). Hirlam is suitable as a reference since it provides a continuous field covering all of Norway at a sub daily time step. In addition, Hirlam gives a higher skill and are less biased than the ENS (Engdahl et al., 2015).

• 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?

AR: Your second suggestion is correct. The raw ensembles from ECMWF (March 2013-Dec 2015) and the calibration parameters (based on data ranging from 2006-2011) were supplied by MetNorway, whereas we did the calibration using the provided calibration parameters and available computer scripts (github/metno/gridpp).

AC: We separate what MetNorway did from what we did. The first paragraph of section 3.1.2 contains the description of calibration parameters from MetNorway, whereas the second paragraphs what we did:

(1) We suggest adding to the first paragraph: "To establish the calibration parameters MetNorway used both ENS re-forecast (Owens, 2018) and Hirlam data from July 2006 to December 2011, both interpolated to a 5×5 km^2 grid… "

(2) And, to the second paragraph: "In this study, we applied the calibration coefficients provided by MetNorway to the temperature forecasts for the period 2013-2015. Accordingly, the ENS was interpolated to the 5×5 km^2 …."

• 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.

AR: MetNorway used the operational Hirlam forecasts for the calibration period. It is correct that Hirlam does not cover the same lead times as ENS. Met Norway established the calibration parameters using the 24 first hours of the forecasts as the reference.
AC: We add a sentence page 7, line 12-13 to clarify this: "The same coefficients, based on the first 24h mapped, are applied to all lead times and ensemble members individually. For forecasts outside the observation range, a 1:1 extrapolation is used. That is, if a forecast is 2°C higher than the highest mapped percentile, then the calibrated forecast is 2°C higher than the same percentile for the reference. "

• By off-setting Tens against Tcal, you create the impression that Tcal is not an ensemble forecast. Consider using Traw and Tcal instead.
AR: We chose to use "ens" instead of "raw", since an elevation-correction was applied the forecasts, and hence they are not actually "raw".
AC: We will clarify in the text that the T_{cal} is an ensemble.

• l29-30. The 'assessment' was done by you, not by the ensemble range.
AR: Thank you. We will rephrase this sentence
AC: Suggestion: "In this study, the sharpness was visually assessed by looking at the ensemble range (i.e. the interval spanned by the lowest and highest forecasted values)"

• 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.
AR: We will plot the empirical distribution of sharpness for all temperature ensembles, and rephrase the sentence concerning sharpness accordingly.
AC: Add a new sentence on sharpness evaluation.

• 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?
AR: We think that the calibration will not affect the spatial and temporal pattern significantly. The calibration function was applied to each ensemble member individually. We therefore kept the order of the ensemble members, both in space and time, and it was not necessary to use the Schaake shuffle.
AC: We think this will be clearer by adding the following description to quantile mapping page 7, line 12-13. (Response above): "… are applied to all lead times and ensemble members individually…"

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.
AR: This is a good suggestion. Nonetheless, we think this is beyond the scope of this study. This could be an interesting topic for a future study.
AC: No change

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?

AR: This is an interesting question. The streamflow forecast at day 5 will be affected by the temperature forecast the previous 4 days as well as day 5. However, for most catchments in this study, the concentration time is less than one day, and the streamflow will respond the same day as a major water input from rain or snow melt. For specific events, it is not evident which of the T-forecasts at day 1-5 is the most important for the Q-forecast at day 5. The sensitivity depends on the sequence of temperature and precipitation. Nevertheless, we think that using temperature CRPSS for day 5 is a good choice since the streamflow at day 5 is the most sensitive to the temperature at day 5 on average (which applies to all lead times). In addition, we see that the improvement in CRPSS across lead times is highly correlated and our results and conclusions would not change if we used temperature CRPSS for days 2, 3, or 4 instead.

AC: Add a sentence in the discussion: e.g. "The result are robust since most catchment in this study have a concentration time of less than one day."

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.

AR: Thank you for highlighting the in-consistency in the use of T_{obs} and T_o. Since the SeNorge temperature is an interpolated product of the observations, we therefore prefer to use T_o.

AC: Changed to T_o in fig 4, and in the text

• L23-25. These sentences are better placed in a discussion section, I think.

AR: OK.

AC: We will move these sentences to the discussion.

• L19 'influence' is missing an 's'. Pls consider replacing by 'affects' though.

AR: Thank you.

AC: We will replacing "influence" with "affect".

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?

AR: We wanted to evaluate the skill of the uncalibrated forecasts as well. If we were to use the uncalibrated as a benchmark, we would not assess the quality of the original forecast, only the change between the uncalibrated and calibrated forecast.

AC: No changes introduced.

Section 5

L7: 'dispersion' is not an expression of quality but a characteristic of an ensemble. Saying 'dispersion improved' makes little sense then?
AR: Thank you. What we mean is that dispersion, as measured by rank histogram convexity, improved.
AC: We change to "Even though both bias and dispersion (i.e. reliability) as measured by rank histogram slope and convexity improved with longer lead time, the reduced sharpness and increased uncertainty, resulted in a reduced skill."

Section 5.1
• L11 "skill. . . depends". Consider replacing by "skill. . . varies with".
AR: Thank you.
AC: We will change as suggested.

• "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.
AR: Good point.
AC: We suggest writing: "Quantile mapping (*as most statistical calibration methods)* is sensitive to forecasts outside the range of calibration values and period (Lafon et al. 2013), this may explain the too high correction in the highest T_{ens} quantile. "

• Immediately following: "and can be a" –> "and *this* can be a"
AR: Noted
AC: Changed as suggested

• 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?
AR: We will clarify that this is a typical situation of winter. This is to some point already exemplified in the text, and we can underline in the text that the situations are typical for winter. (5.3 will be included in 5.1 and 5.2, and we will ensure to get this information in the revised manuscript):
AC: Add "winter" to the existing text: "This seasonal cold bias is also clearly seen in the western catchments Viksvatn and Foennerdalsvatn (Fig. 4). The cold bias in T_{ens} along the coast during winter months can be explained by the radiative land heating and cooling in the coarse resolution forecasts (see Sect.5.1) "

Section 5.2
L10 Grammatically, this sentence is awkward if not wrong.
AR: Thank you; we will rephrase this sentence.
AC: We suggest rephrasing: "These results show that in order to further improve the skill of streamflow forecasts, improved temperature forecasts during the snowmelt season in spring should be in focus. For spring, the streamflow forecasts are sensitive to temperature forecasts. In this study, however, the temperature forecasts were not, for a majority of the catchments, improved by calibration during spring. Thus, we may expect streamflow forecasts to improve if the temperature forecasts themselves are improved."

Figures
Overall
Many figures use a lot of white space between various plots/panels. Consider reducing this or, even better, removing altogether.
AR: We will reduce some white space in figure 1 and 3.
AC: New figures provided

Author Response to RC#3 Jan Verkade

Figure 1

• Do the grey polygons add up to 139 in total? If so, many must be **really** small?

AR: Yes. Especially on the western coast, the catchments are small. This will be clarified in the caption

AC: New caption text: "Figure 1: The maps for Norway shows the 139 catchments used in this study. The left map show the catchment boundaries including the location of four selected catchments. Please note that many catchments are relatively small and difficult to detect. The location of the catchments' gauging stations are shown in the right map. Norway is grouped into five regions (N=north, M=mid, W=west, S=south, and E=east), and all regions are marked with different colors and regional boundaries."

• Caption: consider using 'boundaries' instead of 'limits'

AR: Thank you, we will use 'boundaries'.

AC: See caption text above.

Figure 4

• Why plot the ensemble **mean** and not all five ensemble members, possibly as horizontal lines?

AR: It is not evident to us which modification the reviewer suggests. In this plot, the mean is for the 51 ensemble members not five. If we were to plot all the members, it will be difficult to retain any information. By plotting the mean we show the bias in the forecast and by using the scatter plot, we also show that some biases are dependent on forecasted temperature (a conditional bias).

AC: No changes introduced in the plots.

• 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.

AR: We will unify the axes.

AC: Changed.

Figure 5

• What lead time are these plots for?

AR: Thank you; we will add the lead time in the caption.

AC: Caption update: "All plots are presented for lead time 5 days."

• 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?

AR: We use the same lead time for temperature as for streamflow. See comment to section 4.

AC: No changes applied.

Please consider. . .

• . . . removing data for seasons for which temperature has little or no effect on streamflow levels.

AR: We would like to keep the plots for all seasons here. By showing the difference between the seasons, we think it is easier to understand the large variations we see.

AC: No changes applied

• . . . 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.

AR: We will consider changing the plots. However, unified axes means that we lose information about the regional distribution. An alternative plot with unified axes is presented below.

AC: We prefer to keep the plot as is.

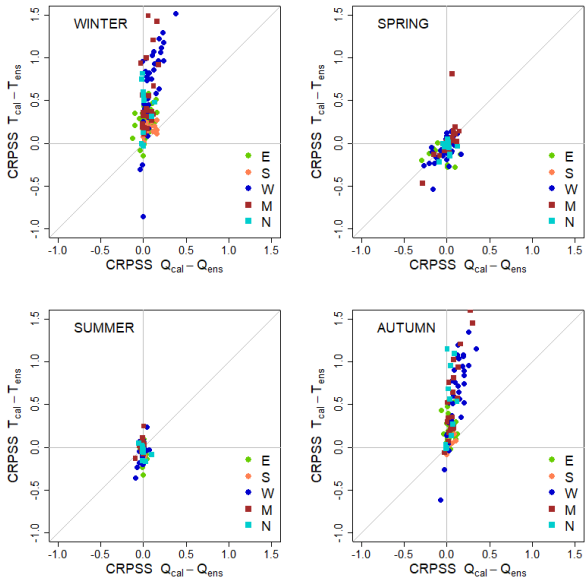


Figure 6

• 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.

AR: We wanted the reader, first of all, to compare CRPSS(T) and CRPSS(Q) Therefore, we placed CRPSS(T) and CRPSS(q) from the spring season on the first line and for the autumn season in the last line. Then the reader can evaluate how the improvements in temperature will affect improvement in streamflow, for both seasons. Secondary, we wanted to show the difference between seasons. Sub-plots for each season are therefore arranged vertically, for both temperature (left) and streamflow (right).

AC: No changes introduced.

• 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)

AR: We prefer to use different scales on the vertical axes within a row to increase the readability of each sub-plot. In particular, the plots of the CRPSS(Q) would be more difficult to read if we used the same scale as in the plots of CRPSS(T) in the left panel.

AC: No changes

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?

AR: Thank you, this will be done

AC: Changed in plot, and in text: p10l28-30: “The horizontal grey dotted lines represent mean annual flood, and the 5-year and 50-year return level for floods in this catchment.”

• 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.

AR: Thank you

AC: We change as suggested

• 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?

Author Response to RC#3 Jan Verkade

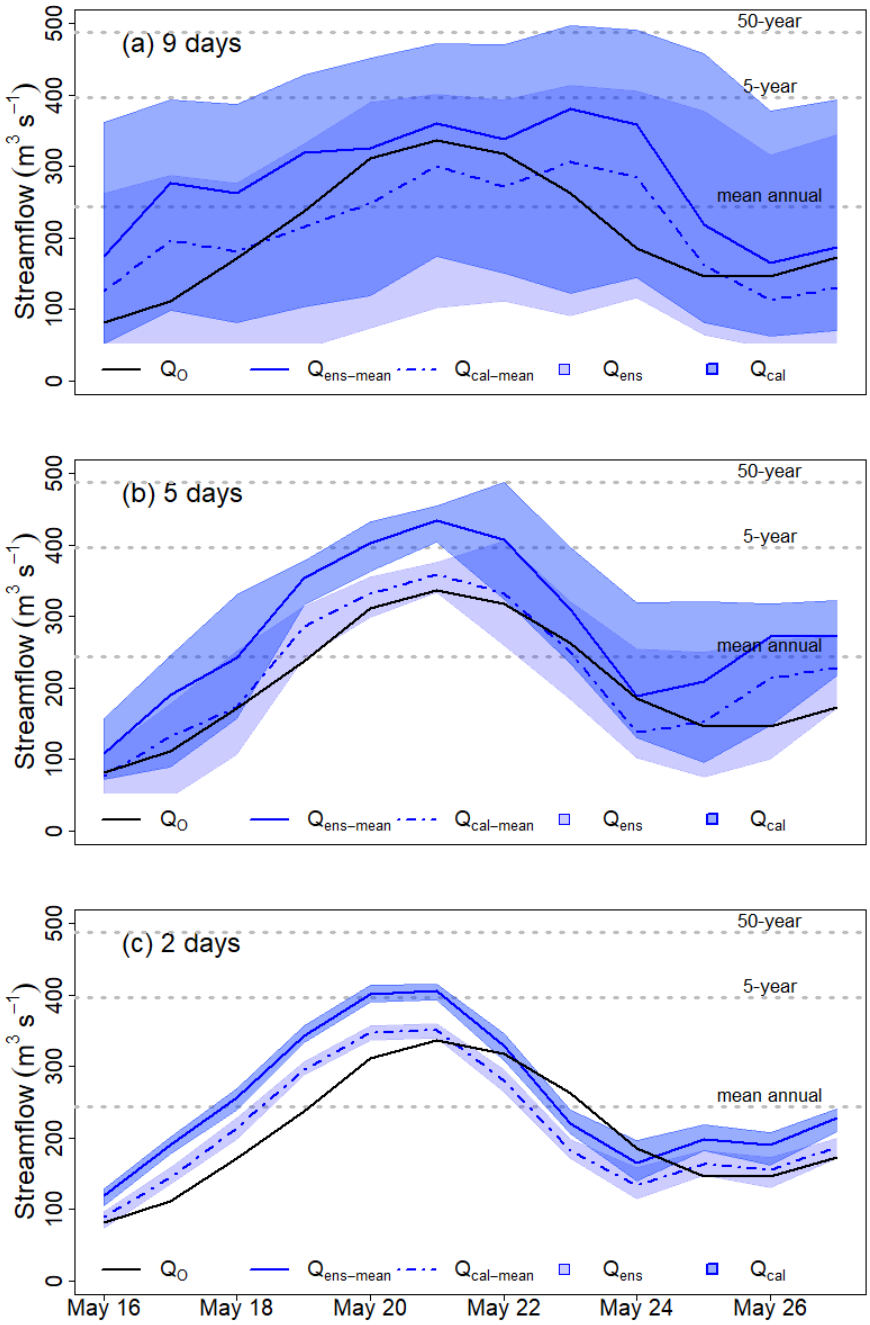
AR: We understand that the introduction of real observations in this figure is confusing, and we will therefore remove the real observation from the figure and from the text.
AC: New plot does not include observation. Need also to rewrite some parts of the text.

• Consider reversing the order of the graphs. The 9d lead time graph was available before the 2d lead time graph?
AR: Thank you, we will change the order of the graphs.

• The horizontal axis labeling is not in English.
AR: Thank you, we will change the labeling to English in al plots
AC: Labeling changed to English

• As all horizontal axes are identical, pls consider removing white space between plots altogether and only label the axis of the bottom plot.
AR: Thank you, this will be done
AC: Label axis only on the bottom, and added (a)-(c) to the different panels. Reduced the amount of white space.

• 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?
AR: In Norway, we use the mean annual, the 5-year and the 50-year floods as exceedance thresholds to issue flood warnings. This figure connects the theoretical aspects to the operational implementation, and points to the importance of calibrated temperature for a flood warning system.
AC: We prefer to keep reference to flood levels, but remove the warning colors all together.



New Reference:

Engdahl, B. J. K and Homleid, M: Verification of Experimental and Operational Weather Prediction Models December 2014 to February 2015. Norwegian Meteorological Institute, METinfo (18/2015), 2015