Supplement:

Especially in hydrometeorological predictions where methods from both the meteorological and the hydrological forecasting community are used, it is of major importance to carefully define the terminology and to coherently use throughout the manuscript.

The current form of the manuscript shows a lack of precise formulations (e.g.: calibration, pre-processing, skill) which should be revised to better communicate the content of the study. Some of the graphics should be enhanced to facilitate the readability and the caption are sometimes incomplete. In addition, more than 15 references mentioned in the text are missing in the reference list and should be added. Furthermore, some additional references could be of interest within discussion to put the findings of the study into a broader picture. Many of the references especially concerning the meteorological forecast are user guides and or technical reports or personal communications, which is fine, but I would appreciate if some more peer-reviewed literature would be cited as there is a large body of existing literature concerning the verification of ECMWF temperature predictions.

In general, the language could be clearer and more concise. To me it is not clear what the authors understand under the term pre-processing, at least in the beginning of the manuscript. E.g. in literature there is a distinction between dynamical and statistical downscaling (see e.g. Li et al. (2017), Yuan et al. (2015)) and statistical downscaling does include a bias-correction. In the present manuscript, the term downscaling does only refer to applying a laps rate correction and interpolation, what is not as downscaling is referred to in the literature. However, I think it would be important for the reader to have a short general overview of what pre-processing is in the introduction. In particular the term calibration, in the present manuscript used as a synonym for bias correction should be introduced more carefully because the term calibration is used by statisticians but in the meteorological, climatological and hydrological communities, the term bias correction is more common.

As you mention the forecasting period used for the study is only two and a half year long which might influence the results. You state this in the discussion but do not explain why it could be critical. I suggest that you discuss this explicitly. Namely, within such a short period, the interannual variability might not be sufficiently covered. In addition, using forecasts from different model Cycles (38r1 to 41r1) might have an influence of the skill as well because the adaption within a new cycle might enhance or decrease the forecast performance making the comparison between seasons difficult as it might not only originate from the particular season but might be influenced by model versions. I suggest including such limitations in the discussion.

To apply Quantile mapping you do need the distribution of the forecast and the distribution of the observations. In section 3.1.2 you state that "MET Norway uses Hirlam temperature forecasts to provide the observational climatology used for parameter estimation". I think here, more information is needed to enable the reader how the calibration is done. Are daily values used for the parameter estimation? Is it empirical or parametric QM used and how are values outside the range treaded (e.g. constant extrapolation)? Is it a member-by-member approach or are the same parameters used for all members? One critical point is that the calibration parameters are interfered from the Hirlam but the hydrological model is run with SeNorge observations. Why are not these observations used? The correction will account for the bias between ECMWF and Hirlam but I would expect that biases with SeNorge will at least slightly differ. Why don't you use the observations from SeNorge to get your calibrations? In the summary it is stated that "The most obvious improvement in the forecasting chain is to use the same temperature information, the SeNorge temperature, for calibrating the temperature forecast that is used for calibrationg the hydrological model, generating …" (P14/L25-27). But if I understand correctly from the manuscript SeNorge and Hrilam are not the same. I have troubles with this procedure as it is known that different forecast models do have different biases. To bias-correct or calibrate ensembles the observations should be taken into account and not another forecast. In this case the bias between two forecasts will be corrected and not the bias of the forecast with regard to the observations. Furthermore, by the many interpolations used, there is a large uncertainty introduced which will lower the trust in the results. Interpolation of ECMEF and Hirlam to derive correction parameters, another interpolation to meet the hydrological model requirements. I think this limitation should be discussed.

Another point that should be discussed is if seasonal correction parameters are really sufficient or does it introduce artificial jumps between periods. In a climate context, seasonal windows for parameter estimation might be sufficient but in an operational forecasting context a shorter window should be taken into account if possible.

In Section 3.2 where the CRPS is introduced you mention different notations (CRPS, S_{crp}) and same for the CRPSS. I think this is confusing, as later in the text only CRPS is used. I suggest only introducing one of the notations and stick to that.

Specific comments:

P1

L7-14: You say the flood forecasting system uses deterministic forecasts for temperature and precipitation). But the ECMWF model you reference provides an ensemble of 51 members. Please state how this is used.

L11-12: "An alternative approach is to use meteorological and hydrological ensemble forecasts" is somewhat misleading. Either you used ensemble meteorological forecasts in combination with hydrological models to generate ensemble streamflow forecasts or one uses a different methodology to produce hydrological ensembles forecasts. I suggest rewriting the sentence: "An alternative approach is to combine meteorological ensemble forecast with hydrological models to quantify the uncertainty in the forecasted streamflow".

L14: "for an accurate forecasting of ", or "to accurately forecast streamflows"

L15: Ensemble forecast of temperature from **the** ECMWF "

L16: "to improve **the** skill and reduce bias**es**"

L18: why do you mention precipitation here? If it is not used for the calibration I would avoid it here.

L20: was used to calculate **the** streamflow

P2

L1: Floods **can** damage… and can have a high …

L5: component**S**

L9: The reference "Müller et al." is missing in reference list

L14: Both reference "Langsrud 1998 a and b" are missing.

L16: as a mean~~s~~ to account for uncertainty in **the** forcing.

L21: The Reference Cloke & Pappenberger, 2009 and Wetterhall et al., 2013 are missing

L25: the ensembles **can be** calibrated

L26: Hamill and Colucci, 1997 and Buizza et al, 2005 are both missing

L29: Gneiting et al. 2005 is missing, Wilks and Hamill 2007 is missing, Raftery et al. 2005 is missing

L30: Evens 2003 is missing

L31: Gneiting et al. 2005 is missing, Wilks and Hamill 2007 is missing. The order of the references is different compared to L29. Bremnes, 2007 is missing.

L31-32: This sentence is very general, it is arbitrary clear that different correction methods do correct the biases differently. I suggest either being more specific about single methods, or to summaries different methods to provide a better overview for the reader instead of listing available techniques. Maybe cite some standard books for statistical bias correction and downscaling (Wilks, 2011) and for forecast verification (Jolliffe & Stephenson, 2011).

P3

L1: snow cover without "–"

L2-4: This sentence is unclear to me. Can you elaborate what you mean?

L5: Gragne, 2015 → missing reference

L7-8: Forecasting, downscaling and interpolation are three completely different things and the challenge is connected to much more than laps rate. For interpolation and downscaling a large part can be attributed to temperature height correction which depend to a large degree to laps rates. But forecasting of temperature is far more complex and related to chaos theory. Rephrase please.

L9: Again, missing references: Aguado and Burt, 2010; Pagès and Miro, 2010, Peter et al., 2010.

L13: Alpine (capital A) as the study looks at catchments in the Alps.

L15: ", **found** only modest….",

L17: I think the effect is not marginal, as you later on show with your results.

L26: do you mean from both, the hydrological and the meteorological perspective?

L27: from **the** ECMWF, in addition I would mention the lead time here but maybe not the MET Norway pre-processing setup as you use the QM to pre-process the forecasts which is, if I understood correctly, not yet part of the pre-processing setup at MET Norway.

L28: Are the retrospective forecasts operational forecast for the period within 2013-2015? This could be misleading for readers or misinterpreted as reforecasts (or hindcasts) which are forecasts for the same day as the operational forecast but for the past 20 years using re-analyses for the initialization. Maybe rephrase to avoid any misinterpretation.

L30: again, I think marginal is the wrong word, if the effect is assumed to be marginal, why should you analyse it in such detail.

L31: Not clear to me. Do you mean that the observed precipitation is used to drive the hydrological model? Specify that to make it clearer.

L33-P4L2: Maybe combine this with the preceding paragraph. This would make it less generic.

P4

L5: **spatial** variations

L6: rather high then steep?

L9: delete "flows"

L18: the smallest catchment has an area of only 3 km^2? Or is it a typo?

L21: what are the selection criteria for "data of sufficient quality"

L27: "og" seems to be Norwegian

L31: snowmelt **driven** flood event

P5

L5: write "available at SeNorge.no"

L7: Mention what kind of interpolation is used (bilinear, kriging, …)

Section 2.2.1:

Mention here that you use the precipitation data from this data set as a substitute of the precipitation forecasts (if this is the case).

L15: constitutes **as** the basis

L20: explain what PEST is.

L21: Abbreviation NS (for Nash-Sutcliffe) not introduced before.

Section 2.2.2:

Is the calibration done for each catchment separately? Do the given values for the NS coefficient represent the mean for all catchments? Is this good? Please state how these values translate into performance compared with other hydrological models.

L22: Missing Reference Gusong (2013), In reference list only Gusong 2016 is listed

2.2.3

To make this more coherent I suggest renaming this section into “Reference observations” (or similar) and in the latter part of the study refer to reference observations as well. Otherwise it is difficult to distinguish between the model stream flow and forecasted streamflow. E.g. on P6 L13 you write reference model run, I assume this is the same as model streamflow? This is somewhat confusing if you state it twice in 2 different paragraphs.

P6

L6: write the lead time as well in days 246 hrs (i.e. 10 days). Why is it 246 hrs and not 240?

L7: The Reference “ECMWF (2018a)” does only provide the documentation and support page of the ECMWF. The Specific documentations can be downloaded. The scientific basis of the ENS system has been discussed in multiple publications and it might be worth to reference some of them and point to this documentation for specific points only.

L8: “the ensemble members of ENS are…”

L9: “with different perturbed conditions **to** represent the …”

3.1

See comment to 2.2.3. I don’t get the difference between model streamflow and reference model run. If I understand correctly these are the same. If so, only describe it in one section. I think here it would be suitable. Reference run = model streamflow, use the same terminology if it is the same.

Are the ENS forecast temporally aggregated as well?

L25: replace “include” with “referred to as”

L27: Use the same units for both grids. ° or km^2. Best would be use both units for both grids, one of them in brackets.

3.1.1

What is the rationale behind the choice of using a nearest neighbour technique?

P7

L4: Bremnes 2007, 2004 are missing in the reference list.

L8: Can you give a reference for the sentence “gives a higher skill and are less biased”

L20: Ensemble forecast verification does not only focus on reliability and sharpness. Therefore, different measures need to be taken into account (as well biases are important).

L30: “lowest and highest forecasted value” does it mean the minimum and maximum? Why not the 10th and 90th percentile and the interquartile range. I think this gives a better estimate of the sharpness of the forecast as it does not only account for the most extreme members.

P8

L12: I would rephrase the sentence. “which a skilful forecast should outperform” and write it in a single sentence.

L18: negative values mean (without s)

L19: “which perform similar to the reference forecast (climatology in this case)”

L20: Do you use here the mean of the daily CRPS? (CRPS with overbar?)

L25-26: This sentence seems to be wrong.

L27: Usually seasons are aggregate in winter = December-February (DJF), spring = mar-may (MAM) and so on. Can you explain your motivation to choose this definition of seasonal aggregation?

P9

L8: as shown in figure…

L9: no comma after “convexity”

The description about the slope and complexity is hard to follow. Could you give an example what the values really tell, e.g. how does a rank histogram look like with a complexity of 2000? I think rank histograms are very useful to be used for visually interpretation and the complexity and slope somehow lead to a reduction of the usefulness of the rank histogram at least to people not familiar with these parameters.

L15: I recommend repeating what T_O and T_{ens} is to enhance the flow in the text.

L27: Same here, mention the abbreviation in brackets in the text to help the reader.

L29: “influenceS” ; Do you mean in streamflow skill or CRPSS?

P10

L4: Do you know why there is no improvement during summer by using calibrated temperatures? Is it due to the absence of snow / snow-melt in summer?

L5: You often mention the Figure number in the last part of the sentence. I personally would prefer this information first what makes it easier to follow the text and figures at once; **It** reveals

L20: What is the significance level you used? I would mention this in the text.

4.3

What are the criteria you used to choose this flood event in May 2016? Mention the motivation for this specific event.

L10: If possible embed this in the floating text and see separate comment to the figure.

P11:

L1: to make it clear I would add: “…increases with lead time (form upper to lower panel).” Linked to my comment on the caption in Figure 10 that it could be misinterpreted as a continuous forecast starting at may 16th.

L4: “**The** box plots … show” (show without s)

5 Discussion

Here I would again use words instead of the Tens Tcal only: “Both raw (Tens) and calibrated (Tens) temperature forecasts were more skilful with …”. I think it makes the text more interesting to read. This could be adapted in different parts of the Manuscript, in the beginning of each section this should be repeated.
L5-9: “Overall, the grid calibration of temperature had a positive effect on both …”, but the lines before it states “…, resulted in reduced skill”. This is somehow contradictive, could you make this clearer?

L18: missing reference Lafon et al. 2013
L20: wrong citation format Ivar Seierstad et al. (2016)
L24: again, wrong citation format Ivar Seierstad et al. (2016)

Subtitles for 5.1 and 5.2 should be coherent “calibration for …” or “calibration for the…”

L26: forecast**s**

P12
L4: “Hence, calculated … “ word at wrong place within sentence.
L7 “indicate” delete additional s

L18: “the bias in T_{ens} is explained by” I think this statement is too strong. It **can** be an explanation, but I think it cannot be reduced to this single causality, as you state in the next sentence.

L21: “The T_{ens} CRPSS is skilful” forecasts have a positive CRPSS and are skilful. The current formulation is not logical, a CRPSS is not skilful.

L28: please state these characteristics very shortly again here.

P13
L1: I don’t understand what you mean with “the averaging effect on temperature skill dominates”.

If I understand correctly, you could discuss here what the difference would be if you use a spatially distributed hydrological model (e.g. gridded version of the model with high resolution). The effect of temperature downscaling might be higher in this case because you do not average temperature again after the downscaling and the spatial distribution within a catchment would have a much larger effect especially in catchments with high spatial variability of soil properties, altitude and vegetation cover.

L13: “the calibrated temperature reduced the skill of the forecasted streamflow.” Please state what skill measure you mean here, did you calculate the CRPSS or bias for that specific event? In the result you only describe the range of the calibrated / uncalibrated ensembles but not a measure of skill.

L15-17: I think you would like to point out that other errors (in the meteorological dataset and the hydrological model) do influence the results. If so, the sentence should be rephrased. Now the reader might think that forecasts are always getting worse if they are calibrated and this would be an argument against your conclusive statement in the summary on Page 14/L19-18.

Figures:

Figure 1: write “grouped” instead of “divided”. Something is wrong in the first sentence “this study shown using”. Please rephrase.

Figure 3:

Avoid overlap of the boxplots to enhance the readability of the plot. There seem to be two line-artefacts on both sides of the figure.

Figure4:

In the text you write T_O and in the Figure it corresponds to T_{obs}. Similarly, T_{ens} and $T_{ens-range.}$ It might facilitate the text of the abbrevations are more consistent in the text, captions and the figures.

Figure 6:

Line artefacts on the left of the figure.

Figure 10:

It is hard to see the actual forecast. I suggest removing the background colors for the warning levels and just plot lines instead. The Figure can easily be misinterpreted as the individual plots (e.g. upper panel for lead day 2) look like a continuous forecast. Maybe it would be more suitable to plot boxplots instead.

Captions: Forecast issue date is the date when the forecast was issued, hence the x-axis could be different for each panel in this figure. I recommend adapting the caption to make this clearer, e.g. target day instead of issue date.

“model streamflow with SeNorge observations” this is $Q_{O.}$ I would write it in brackets as you do for Q_{cal}.

References:

Jolliffe, I. T., & Stephenson, D. B. (2011). *Forecast Verification*. (I. T. Jolliffe & D. B. Stephenson, Eds.), *Forecast Verification*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119960003

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water, 4*(December), e1246. https://doi.org/10.1002/wat2.1246

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. (D. S. Wilks, Ed.) (3rd ed.). London: International Geophysics Series, Vol. 100, Academic Press Inc.

Yuan, X., Wood, E. F., & Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdisciplinary Reviews: Water, 2*(5), 523–536. https://doi.org/10.1002/wat2.1088