Thank you for the positive and thorough evaluation of our article. We appreciate the comments, which are valuable for us 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 response (AR) is marked with red text, while the author's suggestions to corrections (AC) are marked with blue text; we use page and line number to specify the appropriate location, where this is needed. Referee comments are kept in a black text.

Review of 'Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments' by Trine Hegdahl et al.
Anonym referee#4

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.

AR: We thank you for the feedback. We would like to apology for the missing references. It seems to have been an error when we reformatted EndNote, which evidently lead to many references being deleted from the reference list. The main author should nonetheless have detected this flaw prior to submitting the manuscript. We will carefully revise the text to avoid inaccuracies in formulations.

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.

AR: We agree that it is better to use peer-reviewed literature. We chose to use technical reports and personal communication only when necessary and we found no other alternatives. In particular, there are not much peer-reviewed papers on the verification of the ECMWF temperature forecasts for Norway available. Hence, we chose to implement what is available of technical documentation.

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.

AR: We acknowledge that the literature is not consistent in terminology, and particular the terminology differs between the forecasting and the climate projection communities.

In our paper, we chose a terminology that is consistent with a large part of the literature, and that facilitates to explain the approaches we used. We use pre (and post)-processing as a general term, which includes all techniques applied to the raw temperature forecasts in order to improve the temperature output from the atmospheric model (i.e. downscaling and calibration are pre-processing techniques). We pre-processed the temperature in two ways: (i) only downscaling, (ii) both downscaling and calibration, with the purpose to reveal the effect of temperature calibration.

We used the term downscaling on the resampling from a low resolution for the ECMWF forecasts to the 1x1km grid used for the SeNorge data, combined with a temperature correction using a temperature lapse rate. This terminology is used by e.g. UK Met office (Sheridan et al 2010, with references therein). Especially for areas with a complex terrain, where the resolution of the NWP poorly resolves the terrain, the correction for the discrepancy between model elevation and terrain are useful. In some literature, the term downscaling includes both bias correction and resampling, (ref Yuan et al 2015), but we did not use this terminology here.

We used the term calibration on the statistical adjustments of bias and dispersion of the ensembles. The aim of calibration is to make the forecasts reliable in a statistical sense, i.e. 90% of the observations are within a 90% uncertainty interval. In particular, in the meteorological forecasting literature, calibration has this specific meaning (e.g. Gneiting, 2006)

We think that to separate the post-processing into downscaling and calibration is useful, but agree that the term downscaling might have a different signification in parts of the literature. Our terminology is also, to a large degree, in accordance to the descriptions in Li et al (2017). Lie et al (2017) describes the main purposes of post-processing to be the following (1) correct bias and dispersion in the forecasts, (2) to preserve the predictive skill of the forecasts, (3) downscale the forecasts to the scale used in the applications, and (4) to generate ensemble members (…). Further, in the conclusion Li et al (2017) writes that their purpose is "… to calibrate the bias …" In the referred article, we hence see the term calibrate used consistently to describe the statistical properties of both the meteorological and the hydrological ensembles. We further think that using calibration, as part of the pre- and post-processing is a well-established term for the hydrological community using ensemble forecasts. Calibrated ensembles and the calibration methods is more specific than only using only the term pre- or post-processing. Calibration strive for the ensemble to describe the mean and spread of the climatology they should represent.

We have not included any description of the dynamical downscaling, as this usually includes a regional climate model with a different approach, and is not the scope for this study.

AC: We would prefer to keep the notations as is, but we will provide a more detailed description of how we define the term pre-processing, and our use of the terms calibration and downscaling.

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.

AR: We agree that the inter-annual variability might affect the calibration coefficients, and of course, there are aspects with the different model version that might affect the result. However, the changes applied to the different model-cycles did not remove the biases apparent in temperature forecasts (fig 4).

AC: We will address the above mentioned explicit 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?

AR: MetNorway uses parametric quantile mapping based on the first 24h. When a forecast is outside the observation range, a 1 to 1 extrapolation is used. Therefore, if a forecast is 2^{o}C higher than the highest percentile, then the calibrated forecast is 2^{o}C higher than the same percentile for the reference. The same parameters are applied to all members and lead times.

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. I.e. 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. "

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.

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 horizontal 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 used 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 has less (smaller) errors than ECMWF in the temperature forecast for Norway (Engdahl et al. 2015), and as we see from e.g. fig 6 and 7 in this manuscript, the calibration reduces 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 conformation that the Hirlam temperature is less biased than the ECMWF temperature.

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.

AR: We agree that there are uncertainties due to interpolation and downscaling. A temperature calibration that is tailored to the needs for the hydrological modelling would solve this challenge.

AC: We will add a sentence on this in the discussion.

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.

AR: MetNorway provided unique parameters for each month. The parameters are based on a window of three months, which smooths the seasonal patterns. A three month window was chosen to ensure enough data for robust calibration parameters.

In Section 3.2 where the CRPS is introduced you mention different notations (CRPS, Scrp) 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.

AR: We agree that this notation might introduce confusion. The reason is the formatting standard of HESS where equations should only contain one capital letter with sub or super script. However, we find it appropriate to use CRPSS in the text since this is the abbreviation used in the community, and in the equations, we used an alternative notation according to the HESS standard: (S_{crp} and S_{crps} are only used in the equations). This approach is used in many HESS papers.

AC: No change introduced to the manuscript.

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.

AR: The operational system today, uses one deterministic forecast, not the ensemble forecasts. In our setup, the hydrological system is setup to run the 51 ensemble members. We make sure that the same

initial states are used for all members. This is explained in details in the main text, and in the abstract we keep the description simple. We think the suggested changes in the following point also covers this point.
AC: We suggest clarifying this in section 3.1. P6, l24: "In the forecasting mode each temperature ensemble member was used as input and run as separate deterministic forecasts."

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".

AR: You are right. We apply the suggested rewriting.
AC: Applied the suggested rewriting

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 biases"

AR: Thank you. We include the suggestion L14, L15, and L16
AC: We will apply the corrections

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

AR: We mention precipitation since the "observed" precipitation and temperature was used to calculate the initial states of the hydrological model until the forecast issue day. We will consider omitting the sentence about SeNorge in the abstract. Ref RC#3, and discussion on abstract.
AC: We will omit from the abstract: "Estimated observed daily temperature and precipitation were obtained from the SeNorge-dataset, which is station data interpolated to a 1×1 km2 grid covering all of Norway."

L20: was used to calculate the streamflow

AR: Thank you. We include the suggestion
AC: The sentence will be corrected

P2

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

L5: componentS

AR: Thank you.
AC: We will change to the above as suggested.

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

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

AR: Thank you.

AC: We will update the reference list

L16: as a means to account for uncertainty in the forcing.

AR: Thank you.
AC: The sentence will be corrected

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.

AR: All references are now included.
AC: Thank you. We will update the reference list. In addition "Wang and Bishop"

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).

AR: We will cite some standard books and papers that provides reviews of forecast calibration methods.
AC: We will add the following sentence at the end of the paragraph: "A recent review of calibration methods are given in Li et al (2017) and the text book edited by Vanniitsem et al (2018)

- Vannitsem,S. Daniel S. Wilks, Jakob W. Messner, Editor(s): (2018) Statistical Postprocessing of Ensemble Forecasts, Elsevier, ISBN 9780128123720, doi: 10.1016/B978-0-12-812372-0.09988-X.
- 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

P3

L1: snow cover without "–"

AR: Thank you.
AC: We will correct as suggested.

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

AR: We mean that an improvement in temperature forecast will not necessarily translate directly into an improvement of streamflow forecast. If temperatures are well below zero, an improvement in temperature forecasts has no effect on the streamflow forecasts, whereas for temperatures around zero degrees, the streamflow is very sensitive to temperature, in particular when it might turn on or of rain and/or snow melt.

AC: We will rewrite as follows: "The sensitivity of daily streamflow to temperature is non-linear since streamflow depends on temperature thresholds for rain/snow partitioning and for snow melt/freeze processes. The latter depends on the state of the system, i.e. snow is needed to generate snowmelt. For temperatures well below 0°C, the streamflow is not sensitive to temperature, whereas for temperatures around 0°C relatively small changes in temperature might control if the precipitation falls as rain or snow, and consequently, whether streamflow is generated or not."

L5: Gragne, 2015 . missing reference

AR: We will not use this reference in the modified manuscript
AC: The reference will not be used.

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.

AR: You are quite right. We should not have included forecasting in this sentence. We are addressing the downscaling and interpolation of forecasts.
AC: We will remove the word "forecasting" from the sentence.

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

AR: Thanks. We see that in the case of Peter, this is the first name, it should have been Sheridan et al.
AC: All references are corrected and added.

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

L15: ", found only modest….",

AR: Thank you.
AC: We will correct the above as suggested.

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

AR: We used marginal to separate the effect of temperature from that of precipitation. We will change the sentence to 'the isolated effect of…'
AC: We will change to 'the isolated effect of'

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

AR: Yes, we do. This will be clarified in the manuscript.

AC: We change to “Are there spatial patterns in the temperature and streamflow ensemble forecast skill and if so, can these be related to catchment characteristics?”

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.

AR: The information in line 27 is correct. The QM is (was) a part the operational pre-processing chain at MET Norway and is used at the forecast published at yr.no. We chose to not mention lead time here since the choice to focus on lead time 5 days was based on preliminary results.
AC: In section 3.1.2 we add one sentence to clarify: “MET Norway provided temperature grid calibration parameters used in this study. This grid calibration was used in the operational post-processing chain for meteorological forecast including the forecasts published on yr.no.”

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.

AR: We chose retrospective to underline that we used the operational forecasts in retrospect. Nevertheless, we understand that this can be misinterpreted. We will rephrase the sentence.
AC: We will rewrite the sentence as follows: «Three years of operational ECMWF forecasts from 2013-2015 were used to re-generate streamflow forecasts, and the skill of temperature and streamflow forecasts were systematically evaluated for these catchments.

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

AR: OK
AC: We change “marginal” to “isolated”

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.

AR: Yes. The observed precipitation is used to drive the hydrological model.We will rephrase to make clearer
AC: We will rewrite the sentence as follows:
“To investigate the isolated effect of the temperature ensembles on the streamflow forecasts, the observed SeNorge precipitation (Tveito et al., 2005) was used instead of the precipitation ensemble forecasts when we re-generated streamflow forecasts.”

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

AR: We will join the two paragraphs as suggested.
AC: The two paragraphs will be joined.

P4

L5: spatial variations

AR: Thank you.
AC: We will change the sentence as suggested.

L6: rather high then steep?

AR: The Mountains are both high and steep. However, we think that steep is the most important description of the high elevation gradients in the area.
AC: We make no changes in the manuscript.

L9: delete “flows”

AR: Thank you.
AC: The word will be corrected in the manuscript.

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

AR: This is not a typo. There are several small catchments in our dataset, but only one of this size.
AC: there will be no changes in the manuscript

L21: what are the selection criteria for “data of sufficient quality”
AR: This was inaccurate description since the catchments disregarded from the study was due to different reasons, both data retrieving and technical problems. For three catchments, we had problems running the model with the reference data, one catchments there was an issue with the elevation correction, and for two catchments, there were technical problems during the regional analysis. We have a large dataset, so the exclusion of the six catchments will not change our conclusions.
AC: We suggest writing: “Of the 145 flood forecasting catchments, 139 were chosen as the basis for the study (Fig. 1).”

L27: “og” seems to be Norwegian

AR: Thank you.
AC: The word will be corrected.

L31: snowmelt driven flood event

AR: Thank you.
AC: The sentence will be corrected as suggested.

P5

L5: write “available at SeNorge.no”

AR: Thank you.
AC: the sentence will be corrected.

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

AR: The SeNorge temperature is interpolated using kriging on de-trended temperature using standard temperature lapse rates.
AC: We rewrite the sentence to "For this version, gridded temperature is calculated by kriging, where both the elevation and location of temperature stations are accounted for."

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).

AR: Thank you. That is a good suggestion.
AC: We will add a sentence at the end of the paragraph: page 5, line 12
"The SeNorge precipitation substitutes the precipitation forecasts in the ensemble forecasting chain, to reveal the isolated effect of temperature calibration on streamflow forecasts. (see section 3.1 for more details)"

L15: constitutes as the basis

AR: We prefer to keep the sentence as it is.
AC: No changes will be introduced in the manuscript,

L20: explain what PEST is.

AR: We will modify the sentence and explain what PEST is.
AC: The sentence will be rewritten as follows: "... which has been calibrated using the PEST software to establish model parameters (Doherty, 2015) …"

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

AR: Thank you. We will be corrected in the manuscript.
AC: The abbrevation 'NS' will be changed to Nash-Sutcliffe

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.

AR: The calibration is done for each catchment separately. The mean is presented to give an impression of the performance, and of course, there is a great difference in the NS-score between the catchments. We think that NS between 0.73 and 0.77 is ok. Within the range of NS-scores there are of course catchments where the models performs less optimal. Other models applied to the same catchments has a very similar performance, indicating that the quality of data (precipiptation, temperature and streamflow) is an important contribution to model uncertainty. Since we in this paper use the model streamflow in stead of the observed streamflows for evaluation of forecast, we think it is not necessary to provide more details on the calibration of the hydrological model .
AC: No changes introduced in the manuscript.

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

AR: Thank you.
AC: The reference will be corrected to Gusong (2016).

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.

AR: Thank you. We will change “model streamflow” to ”reference streamflow” in the section title and in the text.
AC: We will changed all “model streamflow” to “reference streamflow”.

P6

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

AR: We used lead time 246 hours since we have used the forecast issued at 00:00 aggregated to daily values for the time period 06-06. We can change this to days.
AC: We add one sentence on line 7 to clarify this: “We used the forecast issued at 00:00 and aggregated daily values for the meteorological 24-hour period defined as 06:00-06:00 to provide forecasts for lead times up to nine days.”

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.

AR: We would like to keep the sentence and reference as it is, since this is provides a detailed overview of the model cycles. We provide an additional sentence, including references, to the description of ECMWF.
AC: We add the following sentence to the end of this paragraph: “A more detailed description of the ECMWF ENS system is provided in e.g. Buizza et al. (1999) and Persson (2015).”

L8: “the ensemble members of ENS are…”

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

AR: Thank you.
AC: The sentences will be changed as suggested

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.

AR: We will change to 'model streamflow' to 'reference streamflow', but be prefer to keep section 2.2.3- since we in section 2 describes the data and models, whereas in section 3 we describe how we used the data.
AC: "model streamflow" will be changed to "reference streamflow"

Are the ENS forecast temporally aggregated as well?

AR: The ENS are also temporally aggregated. Ref p7 l1-2 (3.1.1) and l15-16 (3.1.2), and fig 2.
AC: We will add one sentence in line 20: "all temperature forecasts were aggregated to daily values"

L25: replace "include" with "referred to as"

AR: Thank you.
AC: We will change "include" to "refers to"

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.

AR: We think it is more accurate to use use degress for the ECMWF grid, but we will add a parenthesis with the grid resolution in km. Hence, we use degrees and km for EC, only km for SeNorge
AC: We will change as follows: " ... resolution of 0.25° (~ 30km)"

3.1.1

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

AR: We tested also other techniques, e.g. bilinear interpolation, which has a higher computational demand and creates larger output files, than the nearest neighbor interpolation. Since the quality of the forecasts temperature was almost similar, the reduced computing time and smaller storage requirements made the nearest neighbor method more useful.
AC: We introduce no changes in the manuscript.

P7

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

AR: Thank you.
AC: The references will be added.

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

AR: The reference is Engdahl et al 2015
AC: We include this in the text and in the reference list.

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).

AR: In this sentence we refer to a specific paper where the reliability and sharpness is used for evaluation of forecasts. We also think the bias is a part of the evaluation according to reliability. If the forecast is biased it will not be reliable. In the rank-histogram decomposition slope will identify bias in the forecasts.
AC: We introduce no changes in the manuscript.

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.

AR: we agree that specific interquartile range might be a more robust measure for sharpness. Nevertheless, using inter-quantile range does not change the choice we made, and we introduce therefore no changes in the manuscript.
AC: No changes introduced in the manuscript.

P8

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

AR: We think the sentence is fine as it is.
AC: We introduce no changes in the manuscript.

L18: negative values mean (without s)

AR: Thank you.
AC: The word will be corrected as suggested.

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

AR: Thank you.
AC: We will change to "implies that it performs similar to the benchmark (climatology in this case)".

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

AR: Yes, in this case it refers to calculating the average ($\overline{CRPS}$) over all daily CRPS (without an overbar), for the months in question.
AC: No changes will be introduced in the manuscript.

L25-26: This sentence seems to be wrong.

AR: We will reformulate the sentence.
AC: We rephrase: “Finally, we used linear regression to identify relationships between catchment characteristics (here elevation difference and catchment area) and the skill score (T_{cal} and Q_{cal} CRPSS)”

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?

AR: You are right about the usual definition of seasons. We used a different definition since we wanted to isolate a snow melt season, that for most catchment most catchments is in the period April to June. . We think this better seasonal description for streamflow in Norway.

AC: We will add one sentence: “We used this definition of season to better capture a snow melt season that for most catchments in our case study is in the period April to June.”

P9

L8: as shown in figure…

L9: no comma after “convexity”

AR: Thank you.

AC: We will change the sentences as suggested.

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.

AR: We used the convexity and slope since then it is much easier to provide aggregated information of forecast performance. In or results, we do not focus on the values in themselves; the change of the values is the important information. We find that Jolliffe and Primo, 2007 provide detailed information.

AC: We will add one sentence to explain this better:
“As shown in Fig. 3, temperature slope and convexity, improve with increasing lead time, whereas CRPSS gets poorer. The improvement in slope reflects that the under-estimation in the raw ensemble is improved, whereas the improvement in convexity reflects that the under-dispersion is the original forecast is improved. For streamflow, slope gets poorer, convexity improves, whereas CRPSS shows small changes with lead time. “

L15: I recommend repeating what TO and Tens is to enhance the flow in the text.

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

AR: For both comments above, we will repeat the meaning of abbreviations in the beginning of each section.

AC: The sentences will be corrected as suggested.

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

AR: All skill is measured by CRPSS.
AC: Changed to “ affects” Ref RC#3

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?

AR: There are two reasons for the small changes during summer (i) the skill of uncalibrated temperature forecasts are higher in summer and (ii) there is less or no snow in summer, and that will reduce the streamflow sensitivity to temperature.
AC: We will add one sentence: “Two explanations for the small changes in CRPSS during summer are (i) the skill of uncalibrated temperature forecasts are higher and the potential for improvement is lower, and (ii) there is less or no snow in summer, resulting in a reduced streamflow sensitivity to temperature.”

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

AR: We try to vary the placing of the figure number in a sentence and it is a question of style / preference.
AC: We make no specific changes in the manuscript.

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

AR: For the slope of the regression lines being different from zero we used a significance level of p-values < 0,05. This information is available in the caption text for fig. 9. We will consider including this in the text.
AC: We will include the following sentence: “By indicating the significance and sign of the relationships, significant relationships were found for 12 out of 40 regression equations (5% significance level).”

4.3

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

AR: We wanted to present a snowmelt flood event during spring and the selected event in May 2013 in Bulken was a snowmelt flood.
AC: p10, l27: changed from “2014” to “2013”.

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

AR: We are not certain to which line this comment refer.
AC: We added “target days”, to ensure a consistency to figure 10.

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.

AR: Thank you. We will modify the sentence as suggested.
AC: We will change the sentence and add “from lower to upper panel”. Since we have changed the order of panels as suggested in RC#3.

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

AR: Thank you.
AC: The word will be corrected.

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.

AR: We will introduce the abbreviations in the beginning of the sections
AC: We will changed according to suggestion.

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?

AR: The last sentence refer to the difference between raw and calibrated ensembles for all lead times, and we see that the grid calibration improves the performance for most scores and lead times. The previous statements are related to the development of performance for increased lead times. In short, the CRPSS is reduced for increased lead time, it is better for calibrated than raw ensembles.
AC: We will change the sentence: “Overall, the grid calibration of temperature had a positive effect on both temperature and streamflow for most validation scores and lead times.”

L18: missing reference Lafon et al. 2013

L20: L24: wrong citation format Ivar Seierstad et al. (2016)

AR: Thank you.
AC: The reference will be added and the citation will be corrected.

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

AR: Thank you.
AC: The subtitles will be corrected as suggested

L26: forecasts

AR: Thank you.
AC: The word will be corrected.

P12

L4: “Hence, calculated … “ word at wrong place within sentence.

AR: Thank you.
AC: The sentence will be corrected to “Hence, estimated streamflow has a high…”

L7 “indicate” delete additional s

AR: Thank you.
AC: The word will be corrected in the manuscript.

L18: “the bias in Tens 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.

AR: Thank you.
AC: We suggest to replace “is” with “is partly explained by”

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

AR: We will rephrase to clarify that skillful refers to the forecast.
AC: We will change to: “…, CRPSS show that Tens is skillful for both…”

L28: please state these characteristics very shortly again here.

AR: We will modify as suggested.
AC: We will change the text as follows: “ … Only a few significant relationships between catchment charactersitcs, e.g. catchment area and elevation gradient, were found”

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.

AR: What discuss in this paragraph is the effect of catchment size on the performance of the forecasts. We think that a forecast for small catchments are more sensitive than large catchments to the spatial pattern of forecasted temperature. The reasons are that (i) the smallest catchment are smaller than the grid size of the ECMWF model and (ii) it is more challenging to forecast weather on small spatial scales than large spatial scales.
AC: We will replace the last sentence in the discussion: This result is not conclusive, but indicates that (i) the smallest catchment are smaller than the grid size of the ECMWF model and therefore very sensitive to the pre-processing (ii) it is more challenging to forecast weather on small spatial scales than large spatial scales.

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.

AR: You are right. In this sentence, the use of skill is misleading. We did not calculate a specific measure of skill, but merely point to fact that compared to the reference streamflow, the calibrated T forecast induce too high streamflow, and the error becomes larger. A better word would might be performance.
AC: We will changed "skill" to "performance"

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.

AR: We agree, and will add a sentence to clarify this. We will also remove streamflow observations from the figure and consequently from the discussion.
AC: We suggest rewriting: "Firstly, deterioration in the forecast performance using calibrated temperature is particular for this event. Other results provided in this study shows clearly that the calibrated temperature ensembles improve the streamflow ensemble forecasts on average. This discrepancy reveals the other sources of errors; such as the uncertainty of the observed SeNorge precipitation and temperature, and the ability of the hydrological model to capture the highest flood peaks. These points are outside the scope of this study and will not be followed up further here, but are of course important for the performance of a flood forecasting system."

Figures:

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

AR: We will rephrase the caption.
AC: The caption will be rephrased: "The maps for Norway indicates the 139 catchments used in this study. The left map shows the catchment boundaries including the location of four selected catchments.

The right map presents the location of the gauging stations grouped into five regions (N=north, M=mid, W=west, S=south, and E=east), and marked with colors and region boundaries."

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.

AR: We will have a look at the box-plots, the artefacts in the figures will probably disappear in the finishing stage, as all figures will be provided separately. We used partly overlapping boxes for each lead time to increase the readability of the figure, since it is easy to see to which boxes that belongs to the same lead time. We tried without, but found it then more difficult to read the plot.
AC: No changes will be introduced.

Figure4:

In the text you write TO and in the Figure it corresponds to Tobs. Similarly, Tens and Tens-range. It might facilitate the text of the abbrevations are more consistent in the text, captions and the figures.

AR: Thank you. This will be corrected a suggested
AC: We changed the figures, and correct "T_{obs} to T_o" in both plots.

Figure 6:

Line artefacts on the left of the figure.

AR: The artefacts in the figures will probably disappear in the finishing stage; all figures will be provided separately.
AC: We will check that the line artefacts are not present in the final manuscript.

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.

AR: OK. We will do some changes to this figure. Ref A#3. We prefer- however, to not use box-plots. We think that the use of lines and shaded areas increase the readability of the figures.
AC: We have changed the figure. The background removed and the streamflow observation are removed.

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.

AR: .Thank you. We will follow the suggestion.
AC: We will change to "target day" in the caption, and in the text p10-l28

"model streamflow with SeNorge observations" this is QO. I would write it in brackets as you do for Qcal.

AR: Thank you, we will follow the suggestion.
AC: OK. Add both "Obs" and "Q_o" in the text.

References:

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

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