Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments

Trine J. Hegdahl[1,2], Kolbjørn Engeland[1,2], Ingelin Steinsland[3], Lena M. Tallaksen[2]

[1]Norwegian Water Resources and Energy Directorate, Hydrological Modelling, 0301 Oslo, Norway
[2]University of Oslo, Department of Geosciences, 0316 Oslo, Norway
[3]Norwegian University of Science and Technology, Department of Mathematical Sciences, 7034 Trondheim, Norway

Correspondence to: Trine J. Hegdahl (tjh@nve.no)

Abstract. The Norwegian flood forecasting system is based on a flood forecasting model running on catchments located all across Norway. The system relies on deterministic meteorological forecasts and uses an auto-regressive post-processing algorithm to achieve probabilistic streamflow forecasts and thus a measure of uncertainty. An alternative approach is to use meteorological and hydrological ensemble forecasts to quantify the uncertainty in forecasted streamflow. In catchments with seasonal snow cover, snowmelt is an important flood generating process. Hence, high quality air temperature data are important for accurate forecasting of streamflow. In this study, the sensitivity of hydrological ensemble forecasts to the calibration of temperature ensemble forecasts was investigated. Ensemble forecasts of temperature from ECMWF covering a period of nearly three years, from 01.03.2013 to 31.12.2015, were used. To improve skill and reduce bias of the temperature ensembles, the Norwegian Meteorological Institute provided parameters for ensemble calibration. The calibration parameters are derived using a standard quantile mapping method. Estimated observed daily temperature and precipitation were obtained from the SeNorge-dataset, which is station data interpolated to a 1×1 km^2 grid covering all of Norway. The operational flood-forecasting model, a lumped HBV model distributed on 10 elevation zones, was used to calculate streamflow.

The results show that temperature ensemble calibration influenced both temperature and streamflow forecast skill, but differently depending on season and region. We found a close to 1:1 relationship between temperature and streamflow skill change for the spring season, whereas for autumn and winter large temperature skill improvements were not reflected in the streamflow forecasts to the same degree. This can be explained by streamflow being less influenced by sub-zero temperature improvements, which accounted for the biggest temperature biases and corrections during autumn and winter. The skill differs between regions, which could partly be related to elevation differences and catchment area. It is evident, however, that temperature forecasts are important for streamflow forecasts in climates with seasonal snow cover. This indicates that further studies are needed, specifically addressing catchment specific calibration methods, for improved air temperature forecasts.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

1 Introduction

Floods might damage infrastructure, buildings and farmland, and have high economic impacts on society (Dobrovičová et al., 2015). Early warnings based on hydro-meteorological forecasts are an important flood mitigation measure and provide time to reduce flood damage. A flood forecasting system consists of a hydro-meteorological forecasting chain with three main component, all influenced by uncertainties: (i) observations used to establish the initial conditions for the catchment, (ii) meteorological forecasts used as forcing, and (iii) the hydrological model.

The Norwegian flood forecasting system, operated by the Norwegian Water Resources and Energy Directorate (NVE), uses deterministic forecasts of air temperature and precipitation as forcing for hydrological models in 145 catchments across the country. Meteorological forecasts from the AROME-MetCoOp operational weather prediction model (Müller et al., 2017) are used for short range forecasts (day 1 and 2), whereas forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF, 2018a) high resolution model are used for medium range forecasts (day 3 to 9). All forecasts are provided by the Norwegian Meteorological Institute (MET Norway). The Hydrologiska Byråns Vattenbalans model (HBV) (Bergstrom, 1976; Sælthun, 1996; Beldring, 2008) is used as the hydrological forecasting model, which combined with statistical uncertainty models, provides probabilistic streamflow forecasts (Langsrud et al., 1998a; Langsrud et al., 1998b).

An alternative approach to calculate probabilistic streamflow forecasts is to use meteorological ensemble forecasts from numerical weather prediction models as a means to account for uncertainty in forcing. The meteorological ensemble forecasts are created by perturbing the initial states of the original deterministic forecast. The spread of the ensemble members can be interpreted as the uncertainty of the forecasts, where a large spread indicates large uncertainty (Buizza et al., 1999; Persson, 2015). Subsequently, the meteorological ensemble is used as forcing for a hydrological model to produce an ensemble of forecasted streamflow, referred to as a hydrological ensemble prediction system (HEPS). HEPS are increasingly being used in flood forecasting (Cloke and Pappenberger, 2009; Wetterhall et al., 2013). A HEPS adds value to a flood forecast by assessing the forecast uncertainty caused by uncertainties in one or several parts of the modelling chain.

Raw (unprocessed) ensembles are rarely reliable in a statistical sense (Buizza, 1997; Wilson et al., 2007). Reliability means that the observation behaves as if it belongs to the forecast ensemble probability distribution (Leutbecher and Palmer, 2008). To improve reliability, the ensemble forecasts are calibrated by applying statistical techniques correcting bias and under/over-dispersion (Hamill and Colucci, 1997; Buizza et al., 2005; Persson 2015). From a hydrological perspective, pre-processing are techniques (i.e. downscaling and calibration) used on the meteorological ensembles and post-processing refers to techniques used on the hydrological ensembles. Examples of methods used to calibrate meteorological ensembles are ensemble model output statistics (EMOS) (Gneiting et al., 2005; Wilks and Hamill, 2007), Bayesian model averaging (BMA) (Raftery et al., 2005; Wilson et al., 2007), ensemble Kalman filters (Evensen, 2003; Verkade et al., 2013), non-homogenous Gaussian regression (Wilks and Hamill, 2007; Gneiting et al., 2005), quantile mapping (Bremnes, 2007), and kernel dressing (Wang and Bishop, 2005). These methods differ in their sensitivity to length of training data and ensemble size, and how spread and bias are corrected.

In climates with seasonal snow-cover, snowmelt during the spring season is an important flood-generating process. In these climates, temperature is a key variable to classify the precipitation phase and to estimate the snowmelt rate. 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, which may have long-memory effects due to the snow storage (Gragne, 2015). Most Norwegian catchments experience seasonal snow-cover, but are otherwise diverse in terms of the length of the snow season and topographic complexity.

Forecasting, downscaling, and interpolating air temperature in complex topography are challenging, mostly because temperature lapse rates depend on several factors, i.e. altitude, time and place, as well as specific humidity and air temperature (Aguado and Burt, 2010; Pagès and Miró, 2010; Peter et al., 2010). Errors in forecasted temperature might result in a misclassification of precipitation phase and/or cause the hydrological forecasting system either to miss a flood event or provide a false alarm, caused by too high or too low snowmelt rates. It is therefore important to assess the relationship between temperature and streamflow forecasts. The importance of reliable temperature forecasts for streamflow forecasts is demonstrated for two alpine catchments during a heavy precipitation event in Ceppi et al. (2013). An interesting finding in this paper is that catchment elevation distribution, and by this area above the snowline, was important for how streamflow forecasts were affected by temperature uncertainty. Verkade et al. (2013), on the other hand, find only modest effects of temperature calibration on streamflow forecast skill as an average over several years for Rhine catchments.

As far as the authors know, the marginal effect of the uncertainties in temperature forecasts is not yet systematically investigated for a larger number of catchments in a cold climate. The large spatial and seasonal variations in snow accumulation and snowmelt processes found in cold regions with complex terrain require that both spatial and seasonal patterns in the performance of temperature and streamflow forecasts are evaluated.

The main objective of this study is to investigate the effect of temperature forecast calibration on the streamflow ensemble forecasts skill in catchments with seasonal snow cover, and to identify potential improvements in the forecasting chain. In particular, we address the following research question:

- Are there seasonal effects of temperature calibration on the temperature ensemble forecast skill?
- Are there seasonal effects of temperature calibration on the streamflow ensemble forecast skill?
- Are there spatial patterns in the ensemble forecast skill and if so, can these be related to catchment characteristics?

To answer these questions, we applied temperature ensemble forecasts from ECMWF combined with the pre-processing setup from the MET Norway, to 139 catchments in Norway. Three years of retrospective forecasts from 2013-2015 were generated, and the skill of temperature and streamflow forecasts were systematically evaluated for these catchments. In order to investigate the marginal 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. Finally, a flood event is presented, demonstrating the effect of temperature calibration on a single snowmelt induced flood event.

We start by presenting the study area, data and hydrological model (HBV) used (Sect. 2). In Sect. 3, methods used to establish the hydro-meteorological forecasting chain, the skill metrics and evaluation strategy are presented. Section 4 contains the

results, followed by a discussion in Sect. 5. Finally, in Sect. 6, the findings are summarized, conclusions are drawn, and further research questions are discussed.

2 Study area, data and model

2.1 Study area

In Norway there are large variations in climate and topography, and a recent overview over past, current and future climate is given in Hanssen-Bauer et al. (2017). The western coast has steep mountains, high annual precipitation (4000-5000 mm/year) and a temperate oceanic climate. Inland areas have less precipitation, larger differences between winter and summer temperatures, and climatic zones from humid continental, to subarctic and mild tundra (according to the Köpper-Geiger system, see (Peel et al., 2007)). The mean annual runoff flows follows to a large degree the spatial patterns of precipitation. The two basic flood generating processes are snowmelt and rainfall (Vormoor et al., 2015). Most catchments in Norway have prolonged periods of sub-zero temperatures during winter, resulting in a seasonal snow storage, winter low flow, and increased streamflow during spring due to snowmelt. The relative importance of rainfall and snowmelt processes are decided by the duration of the snow accumulation season and the share of annual precipitation stored as snow. Across Norway two basic runoff regimes can be identified, (i) coastal regions with high flows during autumn and winter due to heavy rainfall and (ii) inland regions with high runoff during spring due to snowmelt (Vormoor et al., 2015). However, there are many possible transitions between these two basic patterns (Gottschalk et al., 1979).

The national flood-forecasting system builds on hydrological models providing streamflow forecasts in 145 catchments, covering most parts of Norway, varying in size (~3 to 15447 km^2) and elevation difference (103 to 2284 m). The latter is calculated as the difference between the lowest and highest point on the hypsographic curve, $\Delta H = (H_{100} - H_0)$. The flood forecasting catchments are mostly pristine, although some do have minor (hydropower) regulations. Fourteen catchments have a glacier coverage of 5 % or more. Of the 145 flood forecasting catchments, 139 have data of sufficient quality, and were chosen as the basis for the study (Fig. 1). The catchments were grouped into five ***regions*** based on their location; North (N), South (S), West (W), Mid (M), and East (E) following Hanssen-Bauer et al. (2017) and Vormoor et al. (2016) (Fig. 1, right). These regions are defined by the boundaries of the major watersheds, and reflect major hydro-climatological zones. Rainfall floods dominate in South, West, and Mid, whereas snowmelt floods dominate in East and North. There is still a large variability in hydrological regimes within individual regions. Figure 1 includes the location of four catchments, for which results that are more detailed will be presented. Gjuvaa, Foennerdalsvatn og Viksvatn were used to visualize the challenges in temperature forecasts, and both uncalibrated and calibrated ensemble values will be presented for these three catchments. Viksvatn and Foennerdalsvatn are located in Western Norway and are both catchments with some glaciers (~3 % and 47 % respectively). Gjuvaa is non-glaciered and located in southeast Norway (Fig. 1, left). The Bulken catchment was chosen to demonstrate the effect of temperature calibration on the streamflow forecast for a snowmelt flood event.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

2.2 Observations, hydrological model and forecasts

2.2.1 Interpolated precipitation and temperature observations– SeNorge data

In Norway, a network of about 400 precipitation stations and 240 temperature stations provides daily temperature and precipitation values. These *in situ* observations are interpolated to create a gridded (1×1 km^2) product, referred to as *SeNorge* (SeNorge.no, Tveito et al., 2005). In this study, we used version 1.1. For this version, gridded temperature is calculated by spatial interpolation, where both the elevation and location of temperature stations are accounted for. The observed daily precipitation is corrected for under-catch at the gauges, and triangulation is used for spatial interpolation to a 1×1 km^2 grid. A constant gradient of 10 % per 100 m beneath 1000 meter above sea level (masl) and 5% per 100 m above 1000 masl is applied to account for elevation gradients in precipitation (details can be found in Tveito (2002), Tveito et al. (2005), and Mohr (2008)). The SeNorge data are available from 01.01.1957, and in this study, we used data for the period 01.03.2013 to 31.12.2015 in the forecasting mode and 01.01.1958 to 31.12.2012 to calculate the temperature and streamflow climatology (Sect. 3.2). We hereby denote SeNorge temperature and precipitation, $T_{o[lat,\ lon,\ t]}$ and $P_{o[lat,\ lon,\ t]}$ respectively, where t is an index for observation time. Latitude (*lat*) and longitude (*lon*) represent the grid indexing.

2.2.2 Hydrological model – HBV

The HBV model (Bergstrom, 1976) as presented in Sælthun (1996) and Beldring (2008) constitutes the basis for this study. The vertical structure of the HBV model consists of a snow routine, a soil moisture routine, and a response function that includes a nonlinear reservoir for quick runoff and a linear reservoir for slow runoff. Each catchment is divided into 10 elevation zones, each covering 10% of the total catchment area. Catchment average precipitation and temperature that are elevation adjusted using catchment specific lapse rates were used as forcing. In this study, we used the operational model set-up which has been calibrated using PEST (Doherty, 2015), with Nash-Sutcliffe (Nash and Sutcliffe, 1970) and volume bias as calibration metrics. The calibration, 1996-2012, gives mean NS 0.77, with zero volume bias. The validation period, 1980-1995, shows mean NS 0.73, with a mean volume bias of 5% (personal communication, Gusong (2013))

2.2.3 Model Streamflow

Model streamflow, $Q_{o(c,t)}$, where c is an index for catchment, was derived using SeNorge precipitation and temperature, aggregated to the catchment scale, as forcing to the HBV model (Fig. 2, see “Reference mode” in the green frame). In order to isolate the effect of temperature calibration on forecasted streamflow and avoid effects of hydrological model deficiencies, model streamflow was used as a reference when the streamflow forecasts were evaluated. Similarly, operational flood warning levels (here demonstrated for the case study basin, Bulken), are based on return-periods from model streamflow.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

2.2.4 Temperature ensemble forecasts

We used the ECMWF temperature forecast ensemble (ENS) for the period 01.03.2013 to 31.12.2015 from an original grid resolution of 0.25° (i.e. model cycles/versions 38r1/2, 40r1, and 41r1 (ECMWF, 2018b)). This period covers model cycles/versions for which temperature grid calibration parameters are trained (40r1 and 41r1, see section 3.1.2) plus spring 2013 (cycle 38r1/2) in order to include one more snow melt season. For each issue date d, 51 ensemble members $T_{ens[lat, lon, m, l^*]}$ are provided for a lead time up to 246 hours, where m is the ensemble member and l^* the lead time in 6 hours intervals. The observational time t for a forecast is $d + l^*$. For a full description of the ECMWF ENS product, see ECMWF (2018a). In short, ENS is generated by adding small perturbations to the forecast initial conditions and model physics schemes. Consequently 50 ensemble members are generated running the model with different perturbed conditions, and represents the forecast uncertainty (Buizza et al., 1999; Persson, 2015).

3 Methods

3.1 Ensemble forecasting chain

Figure 2 shows the forecasting modelling chain designed for this study. The green frame presents the observational reference model run that determines the internal states for the forecasting issue date, d, in the red frame. This model run was also used to calculate model streamflow $Q_{o[c,t]}$ (see Sect. 2.2.3). SeNorge temperature and precipitation aggregated to each catchment c ($T_{o[c,t]}$ and $P_{o[c,t]}$) were used to force the hydrological model in the observational reference mode. The red frame illustrates the forecasting mode, including the pre-processing of temperature forecasts. The hydrological ensemble forecasts were calculated using downscaled raw temperature ensemble forecasts ($T_{ens[c,m,l]}$, see Sect. 3.1.1) or downscaled and calibrated temperature ensemble forecasts ($T_{cal[c,m,l]}$, see Sect. 3.1.2), and observed precipitation ($P_{o[c, d+1]}$) as forcing, where m is ensemble member and l is lead time in days. All hydrological forecasts were calculated for all 9 lead times. Note that for each d, the same internal states of the HBV model were used for all ensemble member runs. Thus two sets of streamflow forecasts ($Q_{ens[c,m,l]}$ and $Q_{cal[c,m,l]}$) that differ only by the applied temperature calibration, were derived. The following subsections provide details on the approach used for downscaling and calibration of the ensemble temperature forecasts (ENS).

3.1.1 Temperature forecast downscaling

Within this paper the term downscaling includes the interpolation of temperature from a low resolution grid to a high resolution grid where vertical temperature gradients are accounted for. The ECMWF grid temperature was interpolated from a horizontal resolution of 0.25° to the 1×1 km^2 SeNorge grid, using the nearest neighbour method and aggregated to daily values in order to match the spatial and temporal resolution of the SeNorge data. Due to elevation difference between the ECMWF and SeNorge grid elevations, we corrected the ensemble temperature at the 1×1 km^2 scale by applying a standard atmosphere lapse

rate of -0.65 °C/100 m. Finally, the downscaled temperature ensemble was aggregated to daily values and averaged over the catchment areas to provide $T_{ens[c,m,l]}$ for a given lead time and ensemble member.

3.1.2 Temperature grid calibration

The grid temperature is calibrated using quantile mapping (Seierstad, 2016, Bremnes 2007, 2004) to remove biases by moving the ENS forecast climatology closer to the observed climatology. MET Norway provided temperature grid calibration parameters used in this study. MET Norway uses Hirlam (Bengtsson et al. 2017) temperature forecast (on a 4×4 km^2) to provide the observational climatology used for parameter estimation (calibration). Hirlam is suitable as a reference since it provides a continuous field, gives a higher skill and are less biased than the ENS. To establish the calibration parameters, both ENS re-forecast (Owens, 2018) and Hirlam data from July 2006 to December 2011 are interpolated to a 5×5 km^2 grid. The ENS re-forecast is a 5 member ensemble generated from the same model cycle (40r1 and 41r1) as the ENS. For each grid cell, monthly unique quantile transformation coefficients are determined by using data from a three-month window centred on the target month (e.g. May analysis consists of April, May and June, personal communication (Seierstad, 2017)).
Accordingly, the ENS was interpolated to the 5×5 km^2 grid for which the quantile mapping parameters were used to obtain the calibrated temperature ensembles (T_{cal}). Subsequently, the calibrated ensembles on the 5×5 km^2 grid were downscaled to the 1×1 km^2 grid following the same procedure as for the uncalibrated temperature ensemble (T_{ens}, Sect. 3.1.1). Finally, the calibrated temperature ensemble was aggregated to daily values and averaged over the catchment areas to provide $T_{cal[c,m,l]}$.

3.2 Validation scores and evaluation strategy

The evaluation focused on the performance of the temperature forecast ensembles, and the effect of both uncalibrated and calibrated temperature forecasts on the performance of the streamflow ensembles. A well performing ensemble forecast should be reliable and sharp, where reliability has the first priority (Gneiting et al., 2007). A forecast is considered reliable if it is statistically consistent with the observed uncertainty, i.e. 90% of the observations should verify within the 90% forecast interval. Rank-histograms are often used for visual evaluation of reliability, and show the frequencies of observations amongst ranked ensemble-members. For reliable ensemble forecasts, the rank-histogram will be uniform. A bias in the ensemble forecast is recognized as a slope in the rank-histogram, whereas a U-shape indicates that the ensemble forecast is under-dispersed (Hamill, 2001). In order to quantify the reliability, a decomposition of the chi-square test statistics for the rank-histogram was used to describe the rank-histograms slope (bias) and convexity (dispersion) (Jolliffe and Primo, 2008). Both rank-histogram slope and convexity are negatively oriented with an optimal value of zero for un-biased and uniformly distributed data. The sharpness of a reliable forecast is described by the spread between the ensemble members, where a sharp forecast has a small spread and is the most useful (Hamill, 2007). In this study, the ensemble range (i.e. the interval spanned by the lowest and highest forecasted values) visually assessed the sharpness.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

The continuous rank probability score (CRPS or S_{CRP}) is a summary of reliability, sharpness and uncertainty (Hersbach, 2000). CRPS measures the distance between the observation x_a and the ensemble forecast, where the latter is expressed by the cumulative density function $F_x(x)$:

$$S_{CRP}(F_x, x_a) = \int_{-\infty}^{\infty}[F_x(x) - H(x - x_a)]^2 dx, \quad (1)$$

where H is the Heaviside function that is zero when the argument is less than zero, and one otherwise (Hersbach, 2000). $\overline{CRPS}$ was calculated as the average CRPS over the study period (01.03.2013 to 31.12.2015). $\overline{CRPS}$ is similar to the mean absolute error for deterministic forecasts. The temperature $\overline{CRPS}$ was computed using the SeNorge temperature T_o, as observations, whereas streamflow $\overline{CRPS}$ used $Q_{o[c,t]}$ as observations. This evaluation approach allowed us to evaluate the marginal effect of the uncertainties in the temperature forecasts since we can then, to a large degree, ignore uncertainties in the HBV model itself.

Skill scores are convenient for comparison between forecast variables (e.g. temperature versus streamflow) and catchments since these scores are dimensionless. To calculate the continuous ranked probability skill score (CRPSS or S_{CRPS}), a benchmark score ($\overline{CRPS}_B$ or $\bar{S}_{B\_CRP}$) which a skilful forecast score ($\overline{CRPS}_F$ or $\bar{S}_{F\_CRP}$) should outperform, is needed. For both temperature and streamflow, ensembles representing daily climatology were used as the benchmarks. Daily SeNorge temperature ($T_{o[c,t]}$) from 1958 to 2012 (i.e. 55 years) were used to create a climatological temperature ensemble of 55 members for each day of the year. Similarly, a daily streamflow climatology was established from model streamflow ($Q_{o[c,t]}$) calculated by the HBV model, forced with the 55 years of temperature and precipitation ($T_{o[c,t]}$ and $P_{o[c,t]}$) from the SeNorge data.

CRPSS was calculated for each catchments according to Eq. (2) (Hersbach, 2000).

$$S_{CRPS} = \frac{\bar{S}_{B\_CRP} - \bar{S}_{F\_CRP}}{\bar{S}_{B\_CRP}}, \quad (2)$$

CRPSS varies from -∞ to 1, where one is a perfect score. Negative values means that the forecasts performs worse than climatology, and CRPSS equal to zero implies that it performs similar to climatology. The seasonal skill score was calculated by averaging the daily CRPS only for the months belonging to the target season.

The effect of the grid calibration on the temperature and streamflow forecast skill was evaluated by comparing the validation scores using both the uncalibrated (T_{ens}) and the calibrated (T_{cal}) ensembles to generate the streamflow ensembles.

Spatial patterns in the forecast performance for all 139 catchments, i.e. CRPSS and differences in CRPSS between calibrated and uncalibrated temperature, were mapped for Norway. Further, box plots for the five regions (see Fig. 1) were drawn to reveal potential regional patterns. Finally, we evaluated using linear regression, whether the catchment characteristics, ***elevation difference*** and catchment ***area,*** had any influence on the skill score (T_{cal} and Q_{cal} CRPSS). The linear regression analysis was done for combinations of seasons and regions.

S***easonal*** variations in skill score were assessed by calculating CRPSS for the four seasons winter (January to March), spring (April, to June), summer (July to September) and autumn (October to December).

Finally, the influence of temperature calibration on the flood warning level is illustrated for a snowmelt induced flood event in the Bulken catchment. In the operational flood warning system at NVE, the predefined warning thresholds are catchment specific and calculated return-periods are based on model streamflow, which is also the approach used herein.

4 Results

Temperature and streamflow forecasts were calculated for 139 catchments, 1036 issue dates and 9 lead times. Figure 3 presents a summary of the validation scores, CRPSS and the rank-histogram decomposition, for all lead times. Each box plot shows the variations in the validation scores between the catchments. The rank-histogram slope and convexity describes bias and dispersion in the forecasts, respectively, both can be considered a measure for the reliability. As seen in Fig. 3, temperature slope and convexity, improve with increasing lead time, whereas CRPSS gets poorer. For streamflow, slope gets poorer; convexity improves, whereas CRPSS shows small changes with lead time. To reduce the amount of presented results, the remaining part of this paper focuses on CRPSS for a lead time of 5 days. CRPSS was the chosen validation score since it contains information of reliability, uncertainty and sharpness, and enables a comparison between catchments. A lead time of 5 days was chosen since reliability has improved and some sharpness is maintained.

4.1 Temperature forecasts

Time series of T_o, the range of T_{ens} (left panels) and scatter plots of T_{ens} and T_{cal} versus T_o (right panels) are shown for three selected catchments in Fig. 4. For Gjuvaa, a high altitude catchment in southeastern Norway (Fig. 1), T_o lies within the range of T_{ens} for most days, and temperature forecast T_{cal} was improved by the temperature calibration. The well performing temperature forecasts for this catchment are representative for most catchments in eastern Norway. Representing western Norway, T_{ens} in Viksvatn has a seasonal cold bias that is reduced by the temperature calibration. The cold bias is typical for several catchments in the coastal regions West, Mid and North. Another western catchment, Foennerdalsvatn, has a similar cold bias in T_{ens} to Viksvatn, but for Foennerdalsvatn the bias is notable for all seasons and even increases for T_{cal} (Fig. 4). The Foennerdalsvatn catchment is only 7.1 km^2, has a high elevation, steep topography, 47% is covered by glaciers, and is located close to the coast. The combination of all these catchment characteristics can make forecasting difficult. Foennerdalsvatn is hence an example of how local conditions can be challenging and not well represented, neither by the numerical weather prediction model, nor by the calibration methods.

4.2 Skill – relations to season, spatial location, and catchment characteristics

Scatter plots of the difference between CRPSS for calibrated and uncalibrated forecasts for the temperature and streamflow ensembles are shown in Fig. 5. Each dot represents a catchment and the color indicates the region. The four panels, one for each season, in Fig. 5 shows how the change in temperature CRPSS influence the change in streamflow. For spring, the

relationship is close to the 1:1 line, whereas for winter and autumn streamflow is less sensitive to the temperature calibration. In summer, there are only small changes in CRPSS for both temperature and streamflow. Based on these plots, we chose to present results for autumn and spring for the remaining part of the paper. The summer season was excluded due to the relatively small changes in CRPSS, whereas the winter season was excluded since it performs similarly as the autumn season.

Catchment CRPSS for spring and autumn were sorted according to increasing CRPSS for T_{ens} and Q_{ens} in Fig. 6. The figure reveals that T_{ens} is more skillful in spring than in autumn when T_{ens} has no skill (i.e. CRPSS<0) for about half of the catchments (i.e. they performs poorer than the climatology). In spring, 97% of catchments have skillful temperature forecasts. Temperature calibration improved the temperature skill for most catchments in autumn, whereas for many catchments in spring, the skill worsened. For streamflow, Q_{ens}, there are only small differences in CRPSS between spring and autumn (Fig. 6 right panels). Calibration of temperature improved the skill for streamflow, Q_{cal}, in autumn. Whereas for spring, the streamflow forecast skill followed the temperature skill change, and are both reduced and improved.

CRPSS for uncalibrated temperature and streamflow forecasts, and the change in CRPSS, calculated as the difference in CRPSS between calibrated and uncalibrated forecasts, were mapped for all catchments. Fig. 7 and 8 show the CRPSS values for spring and autumn, respectively. The figures include box plots showing the variations in skill within each region, for both calibrated and uncalibrated forecasts. Neither T_{ens}, nor Q_{ens} skill show any clear spatial pattern in spring (Fig. 7 left panel). For autumn, however, T_{ens} has the lowest skill for the coastal catchments (Fig. 8 left panel). A coastal low CRPSS in autumn is also seen for Q_{ens}, even though less distinct compared to T_{ens}. Both temperature and streamflow CRPSS were improved by calibration for the coastal regions (Fig. 8 right panel).

Table 1 summarizes the result of the linear regression analysis between catchment characteristics (i.e. catchment area and elevation difference) and skill. By indicating the significance and sign of the relationships, significant relationships were found for 12 out of 40 regression equations. Elevation difference is negatively correlated to streamflow CRPSS for the regions East and Mid. Region East also has a negative correlation between streamflow CRPSS and catchment area as opposed to the other regions that have a positive correlation. For none of the regions the correlation changes sign between the seasons. Calibrated temperature and streamflow CRPSS plotted as a function of catchment area are presented for East and South in Fig. 9.

4.3 Snowmelt flood 2013

Forecasts and observations for a snowmelt driven flood in Bulken, Western Norway, are presented in Fig. 10. The figure shows forecasted streamflow for lead times 2, 5 and 9 days for May 16-26 2014. Note that for the lead times 2, 5 and 9 days, the forecasts for e.g. May 18, are issued on May 16, 13 and 9, respectively. The background colors indicate the flood warning levels; green: below mean flood, yellow: between mean and 5-year flood, and orange: between 5-year and 50-year flood. The highest warning level in Norway is red which represents stream flows exceeding the 50-year flood (not shown in Fig. 10). Figure 10 reveals how temperature calibration increases the streamflow for Bulken, leading to a change in warning level for

all lead times. In addition we see how the ensemble spread increases with lead time, from a narrow range around the ensemble mean for the lead time 2 days, to a very wide range for lead time 9 days.

5 Discussion

Box plots of validation scores for all catchments and lead times in Fig. 4 shows that, on average, both T_{ens} and T_{cal} were more skillful with a higher CRPSS, for shorter as compared to longer lead times, and that T_{cal} was more skillful than T_{ens}. Even though both bias and dispersion (i.e. reliability) improved with longer lead time, the reduced sharpness and increased uncertainty, resulted in a reduced skill. For streamflow, the bias increased with longer lead time, while dispersion improved. Further, Q_{cal} was slightly more skillful than Q_{ens}. Overall, the grid calibration of temperature had a positive effect on both temperature and streamflow for all validation scores and lead times.

5.1 Seasonal effect of temperature calibration for the temperature forecast skill

The skill for both T_{ens} and T_{cal} depends on season (Fig. 5 – 8). The relatively small temperature skill improvements in spring and summer, and large skill improvements in autumn and winter, can be explained by the skill of the raw ensembles T_{ens}**.** The low skill for T_{ens} in autumn and winter is caused by the cold bias, and lays the ground for the large improvements seen for T_{cal}. The seasonal differences in skill and response to calibration show the importance of using seasonal calibration parameters. It is also apparent that the applied methods do not perform optimally for all seasons. For spring, the results show that several catchments have a reduction in the forecast skill after calibration. By inspecting the forecasts in detail, we found a too extensive correction of temperature for some days and catchments. Quantile mapping is sensitive to forecasts outside the range of calibration values and period (Lafon et al. 2013), and can be an explanation for too high correction in the highest T_{ens} quantile. The cold autumn bias in temperature forecasts is seen for catchments along the coast (Fig. 8). The relatively low skill and cold bias for sub-zero ECMWF temperature forecasts for the Norwegian coastal areas in cold seasons are documented by Ivar Seierstad et al. (2016). Temperature bias can have complex causes, but the coastal bias typical for Norway is mainly caused by radiation (Hogan et al., 2017). The coarse radiation grid in the ECMWF atmospheric model results in warmer sea points being used to compute longwave fluxes applied over colder land points, causing too much cooling. The radiation resolution is improved in later model cycles (Hogan et al., 2017; Ivar Seierstad et al., 2016).

5.2 Seasonal effect of temperature calibration for streamflow forecast skill

The skill of the temperature calibrated streamflow forecast, Q_{cal}, improved for most of the catchments for autumn, while both improved and reduced skill was seen for spring (Fig. 5 – 8). It is also evident that, independent of the sign of the temperature skill change (Fig. 5), a change in temperature has a larger impact on streamflow in spring than a temperature change has on

streamflow in autumn. An explanation may be that during autumn, for temperatures well below zero degrees, the forecasted streamflow is not affected by improved forecasted temperatures. During spring, temperature are often close to threshold temperatures that control the phase of precipitation and the onset of snowmelt, and such periods are challenging to simulate correctly (Engeland et al., 2010). Calculated streamflow has a hence a high sensitivity to changes in temperature during spring, a sensitivity also described for alpine snow covered catchments by Ceppi et al. (2013). Verkade et al. (2013), on the other hand, found only marginal effects of pre-processing temperature and precipitation for the streamflow skill in the Rhine catchments. The results presented herein and in the cited papers, indicates that the effect of pre-processing depends on the hydrological regime (i.e. sensitivity to temperature), the initial skill of the forcing variables, and for which temporal periods (i.e. for specific events, seasons, or the whole year) the sensitivity is evaluated.

These results show that in order to further improve the skill of streamflow forecasts, temperature forecasts with high skill is most important during the snowmelt season. Forecasts during spring therefore have the highest potential for improvements since the temperature forecasts are not, for a majority of the catchments, improved by the applied calibration. For autumn, the substantial improvement in temperature forecast skill by grid calibration improves streamflow forecast skill, but the sensitivity is less than for spring.

5.3 Spatial patterns

The most pronounced spatial pattern was the low autumn CRPSS for T_{ens} in the coastal areas, also evident from the boxplots for the regions West, Mid and North (Fig. 8). 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 is explained by the radiative land heating and cooling in the coarse resolution forecasts (see Sect. 5.1). In addition, the challenging steep coastal topography is not well represented by the spatial resolution in the ECMWF model (Seierstad et al., 2016). For inland catchments, and the regions south and east, the T_{ens} CRPSS is skillful for both autumn and spring; hence, the calibration has a smaller effect in these catchments.

Autumn streamflow skill was improved by temperature calibration for all regions, the highest improvement was seen for the coast, and the regions West and Mid. From Viksvatn (Fig. 4 panel right) we found that the highest temperature improvements are seen in the temperature range around and below 0 °C. For many coastal catchments, the climate in autumn and winter is partly mild, and temperatures around 0 °C will have an influence on streamflow. In spring, no clear patterns are seen, neither for Q_{ens}, nor for the change in skill.

5.4 Catchment characteristics and skill

Only a few significant relationships between catchment characteristics and skill were found (Table 1). We expected to find the highest temperature skill in large catchments, due to averaging, and in catchments with small elevation differences, due to less elevation correction inaccuracy. No significant relationships between temperature skill and elevation difference was found for any combination of region or season. A positive relationship between temperature skill and catchment area was found for five

out of ten regression equations. This result is not conclusive, but indicates that the averaging effect on temperature skill dominates.

It was expected that streamflow skill would increase with catchment area due to averaging effects. Significant linear regression coefficients were found for east and south but with different signs, the same tendencies for both spring and autumn. The interpretation of this result is therefore ambiguous. For elevation difference, a significant negative correlation was found for three out of ten datasets. This suggest that the downscaling approach has a potential to improve the streamflow forecasts. These results are not conclusive, and studies that are more detailed are needed to determine any significant relationships to catchment characteristics.

5.5 Snowmelt flood 2013

The snowmelt flood event (Fig. 10) illustrates clearly how temperature calibration influences forecasted streamflow. The increase in forecasted temperature by grid calibration, results in additional snowmelt and thus increased streamflow. The increased streamflow led to a change in the warning level from green/yellow to yellow/red, and moves closer to the observations. For this event, however, the use of calibrated temperature reduced the skill of the forecasted streamflow, Q_{cal}. The model streamflow, Q_o that is used as the reference, is better captured by the streamflow forecasts based on uncalibrated temperature forecasts, Q_{ens}. 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.

The ensemble range for the snowmelt event clearly increases with increasing lead time. For a lead time of 2 days (upper panel) the range is too narrow, while for a lead time of 9 days (lower panel), the wide forecasting intervals capture the events, but there is little information left in the forecasts.

6 Summary and conclusion

The main objective of this study was to investigate the effect of temperature forecast calibration on the streamflow ensemble forecast skill, and to identify potential improvements in the forecasting chain. We applied a gridded temperature calibration method, and evaluated its influence on both temperature and streamflow forecasting skill. The seasonality in skill was evaluated and correlations to catchment characteristics and spatial patterns were investigated. Supported by the results presented in this paper, our answers to the research questions listed in the introduction are summarized as follows:

Are there seasonal effects of temperature calibration on the temperature ensemble forecast skill?

- The largest temperature skill improvements by calibration were found for low performing coastal catchments in autumn and winter.

- The effect of calibration on temperature skill was less clear in spring and summer. In spring, the calibrated temperature resulted in reduced skill for many catchments.
- Smaller bias in spring and summer explained a higher T_{ens} skill and hence, less room for improvements by calibration.

Are there seasonal effects of temperature calibration on the streamflow ensemble forecast skill?

- In autumn and winter, streamflow skill improved for most catchments. For spring, the calibration resulted in both better and worse skill.
- In spring, changes in temperature skill had a higher influence on streamflow skill, compared to autumn and winter. Summer showed small changes for both temperature and streamflow.

Are there spatial patterns in the ensemble forecast skill and if so, can these be related to catchment characteristics?

- The skill in temperature forecasts was the lowest in coastal catchment along the coast in West, Mid and North in autumn, caused by a cold bias in the forecasts (this was also the case for winter, although these results are not shown).
- The largest improvement in skill for both temperature and streamflow was found for catchments with a cold bias in the temperature forecasts.
- A regional division seemed useful to identify spatial patterns in temperature forecasts, whereas for streamflow the spatial patterns were not so obvious.
- It was not possible to conclude a relationship between the catchment characteristics and skill.

Snowmelt flood

- Streamflow increased by temperature calibration, changing the flood warning level, clearly showing the importance of correct temperature calibration for catchments with snow during snowmelt season

This study showed that the applied gridded temperature calibration method improved the temperature skill for most catchments in autumn and winter. Temperature forecasts have an impact on streamflow, and are important for seasons where temperature determines snowmelt and discriminates between rain and snowfall. The improvement in temperature skill propagated to streamflow skill for some, but not all, catchments. This was to a large degree depending on region, and the skill of the uncalibrated ensemble.

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 calibrating the hydrological model, generating the initial conditions for the hydrological system, and evaluating the performance. In particular, the calibrated temperature forecast could be improved during spring when the streamflow forecasts are the most sensitive to temperature. The pre-processing of temperature includes both the elevation correction dependency on lapse rate and the calibration method. Lapse rate in this study is defined as a constant, but actually depends on weather conditions, location and elevation. In addition, the calibration method, here the quantile mapping, is sensitive to forecasted values outside the observation range, and other methods should be considered.

The conclusions in this study are based on a testing period of almost three years. Even if this is a relatively short testing periods, we believe that the large number of catchments to a large degree compensates for the short testing period and that the results and conclusions are therefore relatively robust.

The conclusions herein are based on a large and relatively representative data-set from Norway, but we suggest that some of the main conclusions can be valid for regions with a similar climate. The most important general conclusion is that streamflow forecasts are sensitive to the skill of temperature forecasts, especially in the snow melt season. In addition, this study shows that reducing the cold temperature bias in in coastal areas results in improved streamflow forecasts, and that the pre-processing need to account for seasonal differences in temperature forecasts (biases).

7 Data

Processed data is available by contacting corresponding author. Raw meteorological data must be required directly from ECMWF.

8 Authors contribution

T. J. Hegdahl prepared the data, set up the forecasting chain (including writing new code for non-available functionalities), did the data simulations and analysis, and wrote the manuscript. K. Engeland contributed in the writing. K. Engeland, I. Steinsland and L. M. Tallaksen contributed in the design of the study, by advice during the work, and in the revision of the manuscript.

9 Acknowledgement

The authors would like to thank B. Grønbech at NVE for the work done with setting up the hydrological model for ensemble forecasting. We would also like to thank A. Singleton at MET Norway for his comments during the work and proofreading of the manuscript. T. Nipen and I. Seierstad at MET Norway we thank for their support and sharing of precipitation and temperature ensemble forecast calibration knowledge. In addition, their aid was valuable during the implementation of https://github.com/metno/gridpp in the forecasting chain. Thanks also to colleagues at NVE working in the project “Better uncertainty estimation in flood forecasting”, led by E. Langsholt.

10 References

Beldring, S.: Distributed Element Water Balance Model System.Norwegian Water Resources and Energy directorate, report 4, 40 pp, Oslo, 2008.

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., de Rooy, W., Gleeson, E, Hansen-Sass, B., Homleid, M., Hortal, M., Ivarsson, K-I., Lenderink, G., Niemelä, S., Nielsen, K. P., Onvlee, J., Rontu, L., Samuelsson, P., Muñoz, D.S., Subias, A., Tijm, S., Toll, V., Yang, X., and Køltzow, M. Ø.: The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System. Monthly Weather Review, 145(5), 1919-1935. doi:10.1175/mwr-d-16-0417.1, 2017.

Bergstrom, S.: Development and application of a conceptual runoff model for Scandinavian catchments.Swedish Meteorological and Hydrological Institute, 1976.

Buizza, R., Milleer, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 125(560), 2887-2908. doi:10.1002/qj.49712556006, 1999.

Ceppi, A., Ravazzani, G., Salandin, A., Rabuffetti, D., Montani, A., Borgonovo, E., and Mancini, M.: Effects of temperature on flood forecasting: analysis of an operative case study in Alpine basins. Natural Hazards and Earth System Sciences, 13(4), 1051., 2013.

Doherty, J.: Calibration and Uncertainty Analysis for Complex Environmental Models. Brisbane, Australia, 2015.

ECMWF. Set III - Atmospheric model Ensemble 15-day forecast (ENS). Retrieved from https://www.ecmwf.int/en/forecasts/datasets/set-iii, 2018a.

ECMWF. Changes in ECMWF models. Retrieved from https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, 2018b.

Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model. Journal of Hydrology, 384(1), 142-155, 2010.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 243-268. doi:10.1111/j.1467-9868.2007.00587.x, 2007.

Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R., and Tollan, A.: Hydrologic Regions in the Nordic Countries. Hydrology Research, 10(5), 273-286, 1979.

Gusong, R.: personal comment 15.06.2016 [Calibration of HBV - NVE flood forcasting], 2016.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review, 129(3), 550-560, 2001.

Hamill, T. M.: Comments on "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging". Monthly Weather Review, 135(12), 4226-4230. doi:10.1175/2007mwr1963.1, 2007.

Hanssen-Bauer, I., Førland, E. J., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J.E.Ø., Sandven, S., Sandø, A.B., and Sorteberg, A.: Climate in Norway 2100 - a knowledge base for climate adaption. Tech. Rep. 1, Norwegian Climate Service Centre, 2017.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting, 15(5), 559-570. doi:10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

Hogan, R. J., Alhgrimm, M., Balsamo, G., Beljaars, A. C. M., Berrisford, P., Bozzo, A., Di Giuseppe, F., Forebes, R.M., Haiden, T., Lang, S., Mayer, M., Polichtchouk, I., Sandu, I., Vitart, F., and Wedi, N.: Radiation in numerical weather prediction Technical Memorandum: ECMWF. Retrieved from: https://www.ecmwf.int/en/elibrary/17771-radiation-numerical-weather-prediction., 2017.

Jolliffe, I. T., and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic. Monthly Weather Review, 136(6), 2133-2139, 2008.

Mohr, M.: New routines for gridding of temperature and precipitation observations for "SeNorge. no". Met. no Report, 8, 2008.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles. Journal of Hydrology, 10(3), 282-290. doi:10.1016/0022-1694(70)90255-6, 1970.

Owens, R. G., and Hewson, T. D.: ECMWF Forecast User Guide. Retrieved from Reading: https://confluence.ecmwf.int/display/FUG/Re-forecasts, 2018.

Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification. Hydrology and earth system sciences discussions, 4(2), 439-473, 2007.

Persson, A.: User guide to ECMWF forecast products. In E. Andersson & I. Tsonevsky (Eds.), Reading, 2015.

Seierstad, I.: personal comment 10.11.2017 [Temperature calibration parameters], 2017.

Seierstad, I., Kristiansen, J., and Nipen, T: Better temperature forecasts along the Norwegain coast. Newsletter. 148- Summer 2016. Retrieved from https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast, 2016.

Sælthun, N. R.: The Nordic HBV model. Norwegian Water Resources and Energy Administration Publication, 7, 1-26, 1996.

Tveito, O. E.: Spatial distribution of winter temperatures in Norway related to topography and large-scale atmospheric circulation. Paper presented at the IAHS PUB Workshop, Brasilia, Brazil, 2002.

Tveito, O. E., Bjørdal, I., Skjelvåg, A. O., and Aune, B.: A GIS-based agro-ecological decision system based on gridded climatology. Meteorological Applications, 12(1), 57-68. doi:10.1017/S1350482705001490, 2005.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology, 501, 73-91. doi:https://doi.org/10.1016/j.jhydrol.2013.07.039, 2013.

Vormoor, K., Lawrence, D., Heistermann, M., and Bronstert, A.: Climate change impacts on the seasonality and generation processes of floods - projections and uncertainties for catchments with mixed snowmelt/rainfall regimes. Hydrol. Earth Syst. Sci., 19(2), 913-931. doi:10.5194/hess-19-913-2015, 2015.

Vormoor, K., Lawrence, D., Schlichting, L., Wilson, D., and Wong, W. K.: Evidence for changes in the magnitude and frequency of observed rainfall vs. snowmelt driven floods in Norway. Journal of Hydrology, 538, 33-48, 2016.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

Hydrology and Earth System Sciences
Discussions

Table 1: Summary of significant correlations between CRPSS for calibrated temperature (T_{cal}) and streamflow (Q_{cal}) ensembles and catchment characteristics, i.e., area and elevation difference (ΔH), for the five regions. Blue color indicates a significant positive relationship, red a significant negative relationship, and grey a non-significant relationship. Results are for a lead time of 5 days.

5

		T_{cal}	Q_{cal}	T_{cal}	Q_{cal}
		SPRING		AUTUMN	
Area (km^2)	East				
	South				
	West				
	Mid				
	North				
ΔH (m)	East				
	South				
	West				
	Mid				
	North				

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

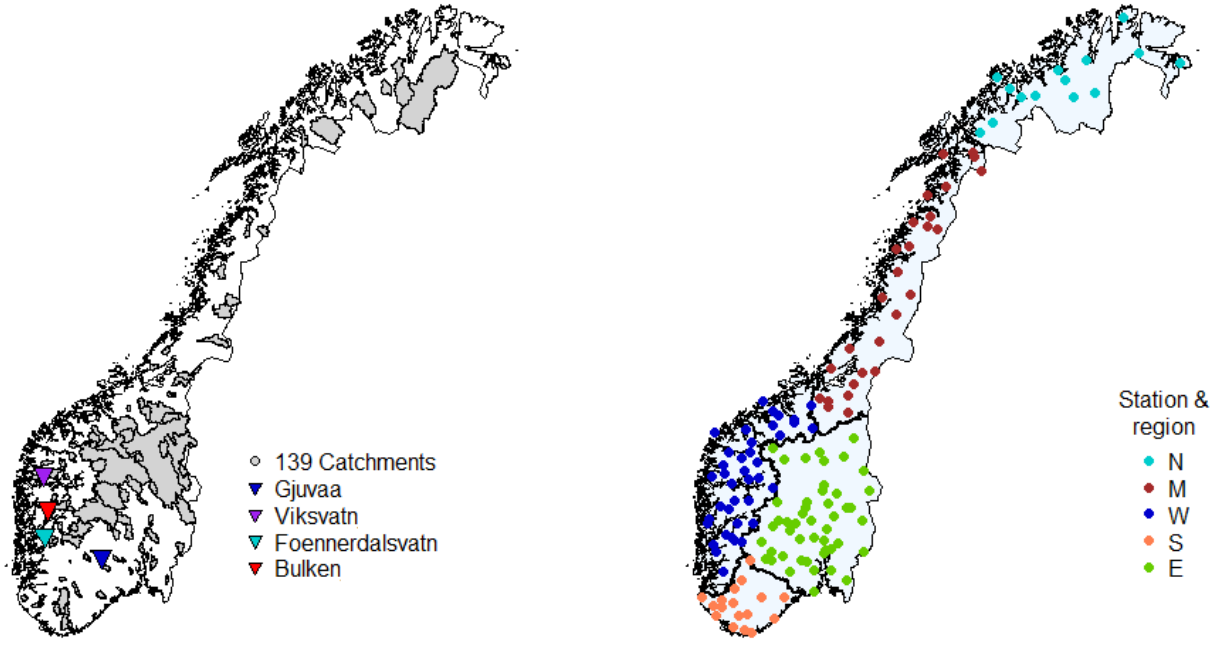

Figure 1: Norway with the 139 flood forecasting catchments used in this study shown using catchment boundaries (left map) and the location of the gauging stations (right map). The location of the four example catchments are shown in the left map. The catchments are divided into five regions (N=north, M=mid, W=west, S=south, and E=east) marked with colors and regional limits in the right map.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

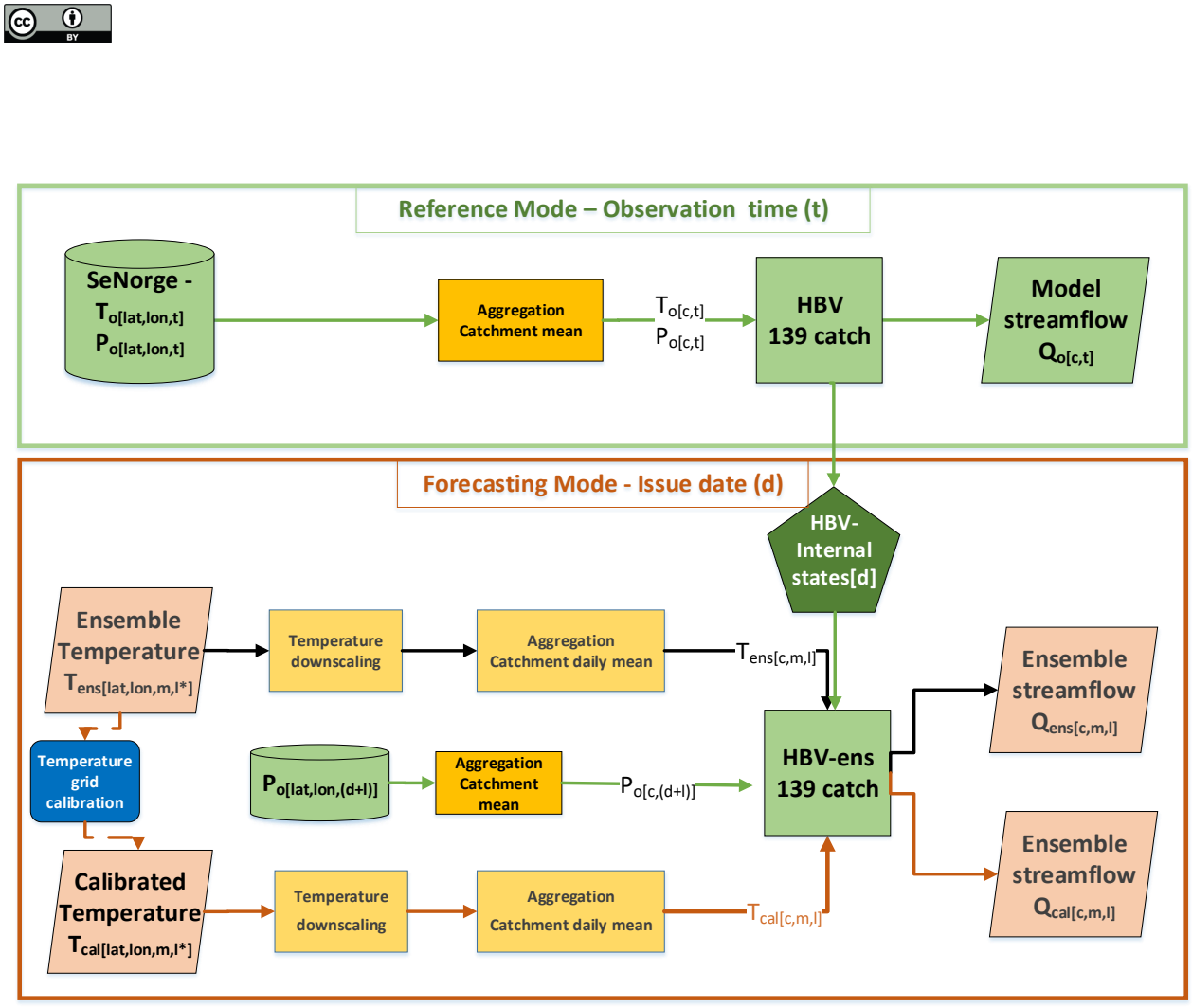

Figure 2: Conceptual diagram of the ensemble forecasting chain. The upper green frame shows the reference mode that is the calculation of modelled streamflow using the HBV model with catchment aggregated daily mean values of SeNorge temperature (T_o) and precipitation (P_o). In the forecasting mode, the lower red frame, ECMWF temperature ensembles are downscaled to 1×1 km^2 prior to catchment aggregation. Calibrated temperature (T_{cal}) is calculated from T_{ens}, applying a grid calibration at 5×5 km resolution. Daily average forecast values (T_{ens} or T_{cal}) and observed precipitation (P_o) are used to force the hydrological model at forecasting issue date (*d*), with internal states from the reference mode.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

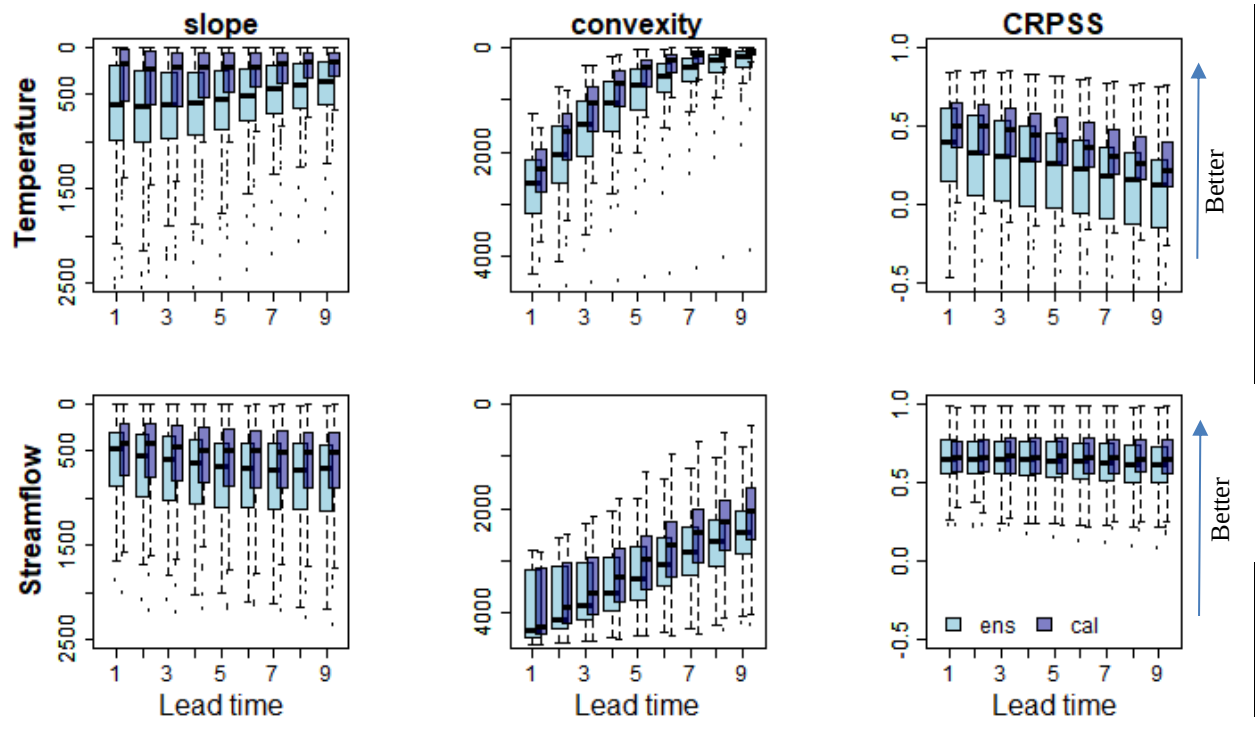

Figure 3: Summary of temperature and streamflow scores for all lead times. Each box represents the 139 catchments values averaged over all issue dates. Rank-histogram test decomposition for slope and convexity is shown in first and second column respectively and CRPSS in the third column. Temperature in the top row and streamflow in the bottom row. Results are based on the full dataset, and are shown for both uncalibrated (light blue) and calibrated (blue) ensembles at lead times 1 to 9 days. For slope and convexity, zero is the optimal value, and the scales are reversed so that the optimal value is on the top, corresponding to CRPSS optimal value at 1.0.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

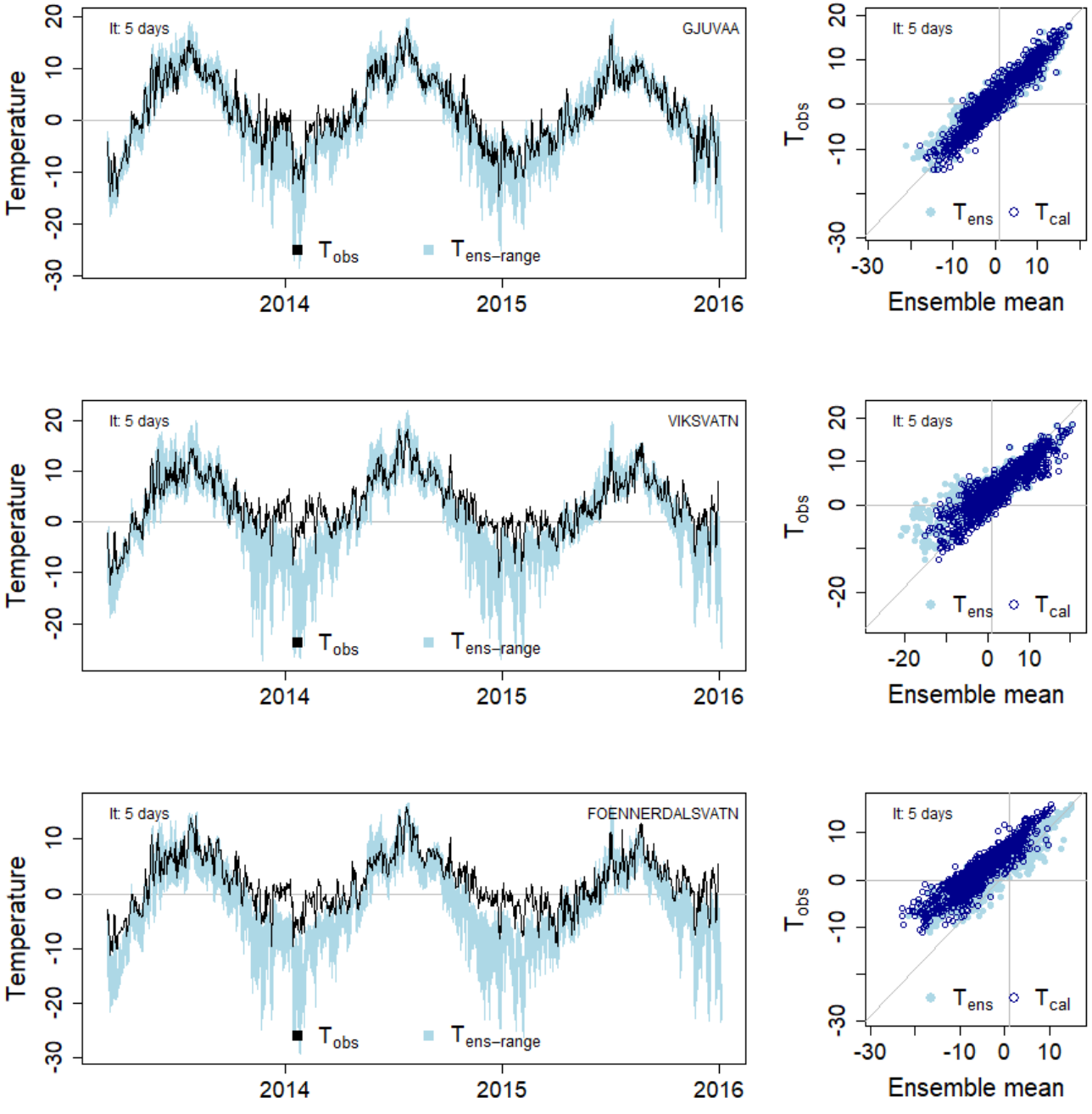

Figure 4: Timeseries of temperature for Gjuvaa (upper), Viksvatn (middle) and Foennerdalsvatn (lower) showing the range of uncalibrated temperature ensemble forecast ($T_{ens-range}$, lightblue area) for the period 2013-2015, SeNorge observations are shown as black lines. Scatter plots show ensemble mean temperature for both calibrated (T_{cal}, blue) and uncalibrated (T_{ens}, lightblue) temperature plotted against SeNorge temperature (T_{obs}). Lead time is 5 days.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

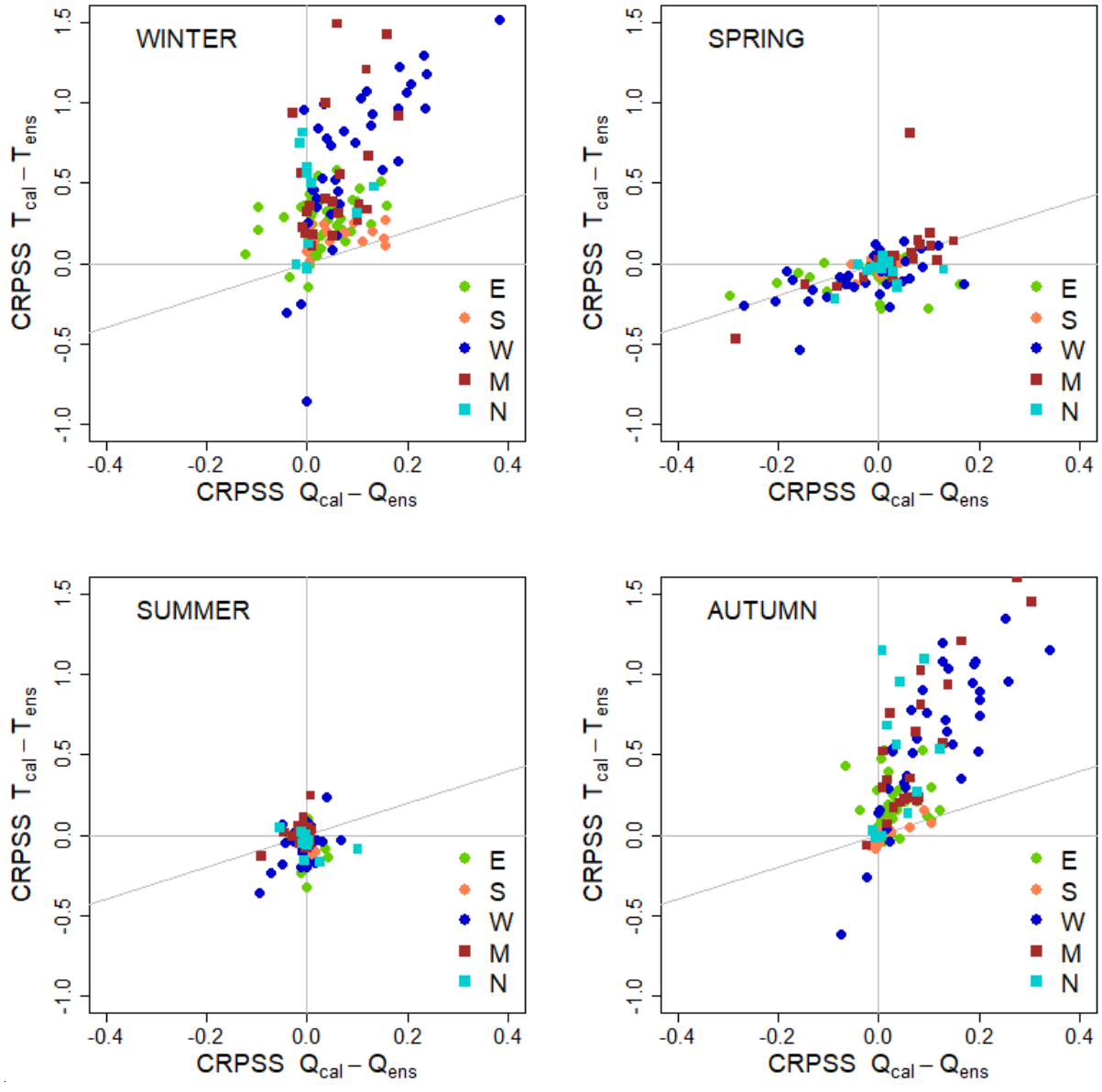

Figure 5: Difference in CRPSS for uncalibrated and calibrated temperature for the four seasons. Difference in temperature skill plotted on the y-axis and difference in streamflow skill on the x-axis. Catchment values are color indexed by region.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

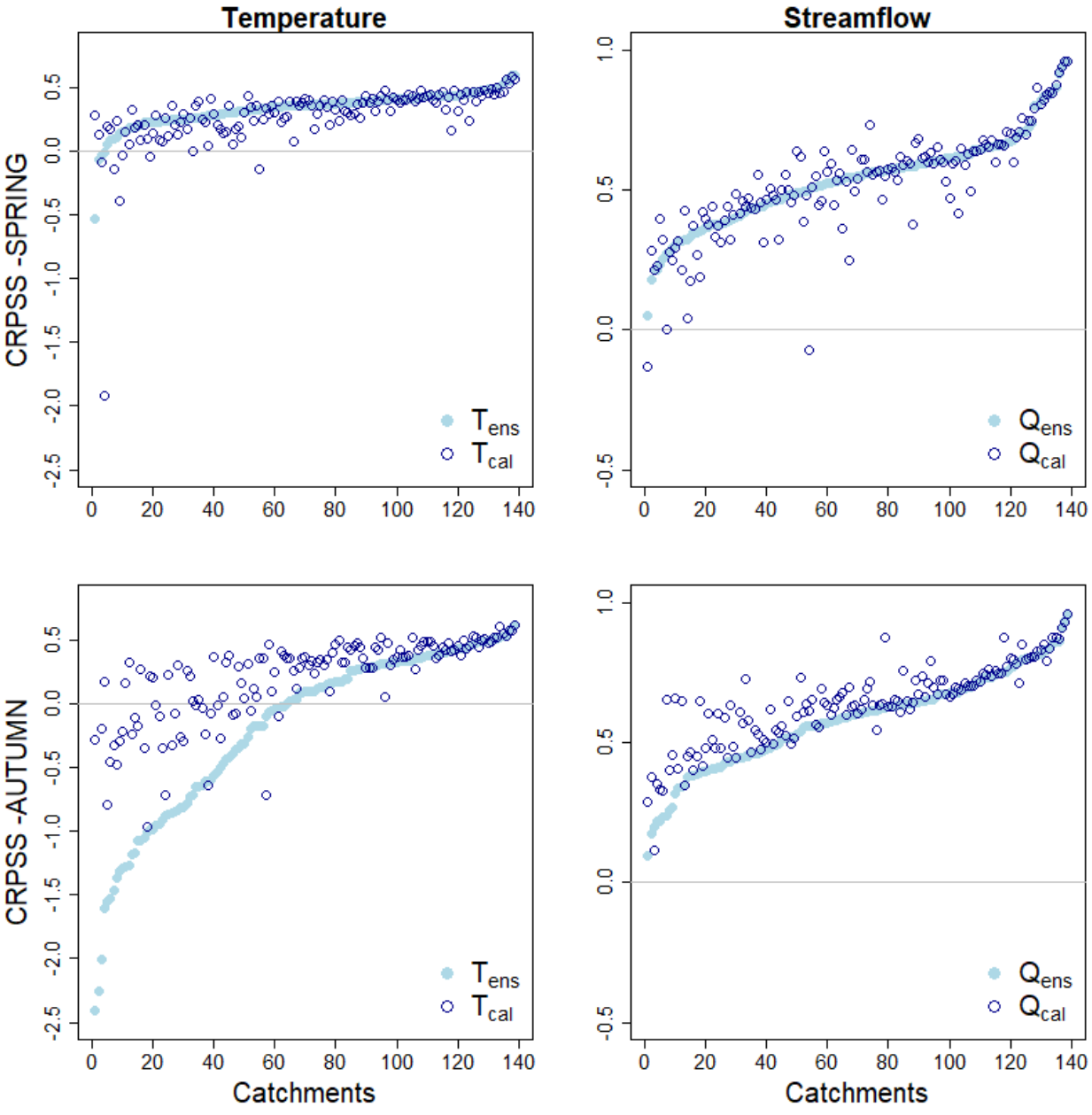

Figure 6: Temperature (T_{ens} and T_{cal}, first column) and streamflow (Q_{ens} and Q_{cal}, second column) CRPSS for SPRING (top) and AUTUMN (bottom). The catchments are ordered by increasing CRPSS for T_{ens} and Q_{ens} (light blue dots), the catchment calibrated values (T_{cal} and Q_{cal}) are plotted as blue circles. All results are presented for lead time 5 days.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

Hydrology and Earth System Sciences
Discussions

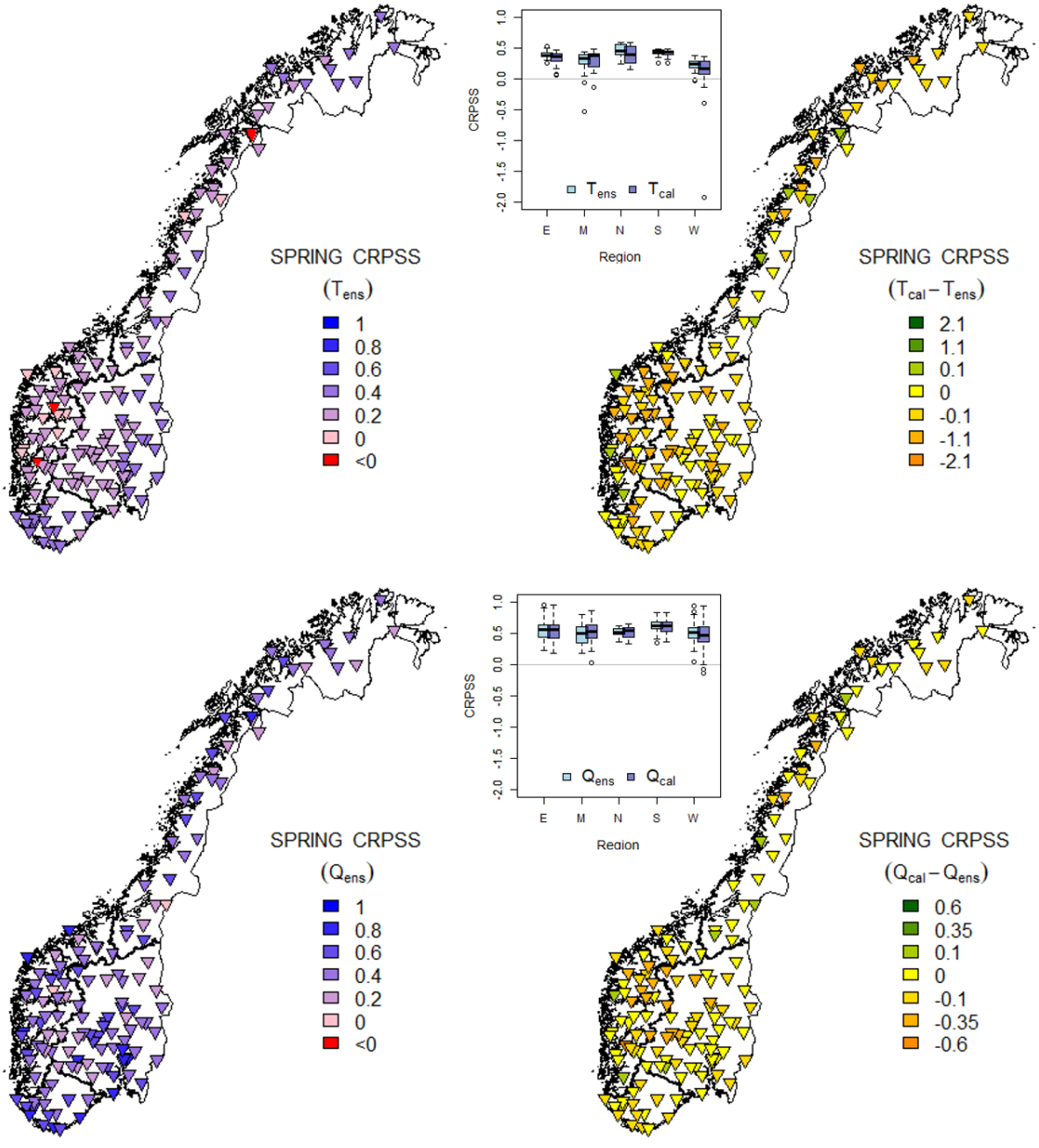

Figure 7: Spring CRPSS for uncalibrated forecasts (left maps) and CRPSS difference between calibrated and uncalibrated forecasts (right maps) for temperature (upper panel) and streamflow (lower panel). A darker blue color (left maps) indicates an optimal performance (maximum CRPSS=1.0), pink a CRPSS of zero, and red a negative value. A green color (right maps) indicates a positive effect of temperature calibration on the skill, yellow means no effect, and orange color indicates a negative effect. The boxplots show temperature and streamflow CRPSSs grouped by region (Fig. 1). All results are presented for lead time 5 days.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

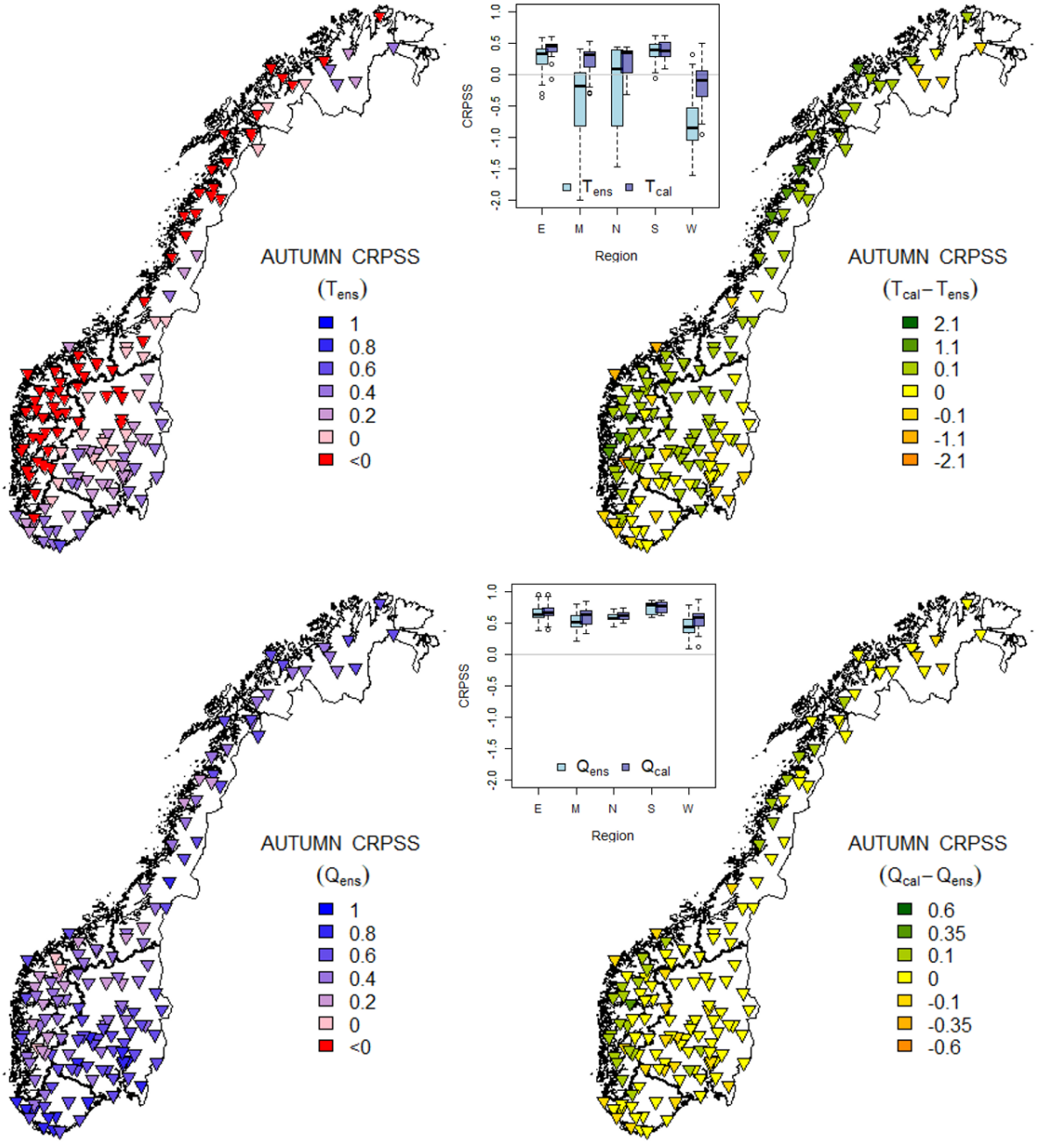

Figure 8: Autumn CRPSS for uncalibrated forecasts are presented to the left where darker blue color indicates an optimal performance (maximum CRPSS=1.0), pink color represents a CRPSS of zero, and red negative. The difference in CRPSS between calibrated and uncalibrated forecasts are presented to the right, where green color indicates a positive effect of temperature calibration on the skill, yellow zero, and orange color indicates a negative effect. The boxplots of both calibrated and uncalibrated temperature and streamflow CRPSS show catchments grouped by region (Fig. 1). All results are presented for lead time 5 days.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

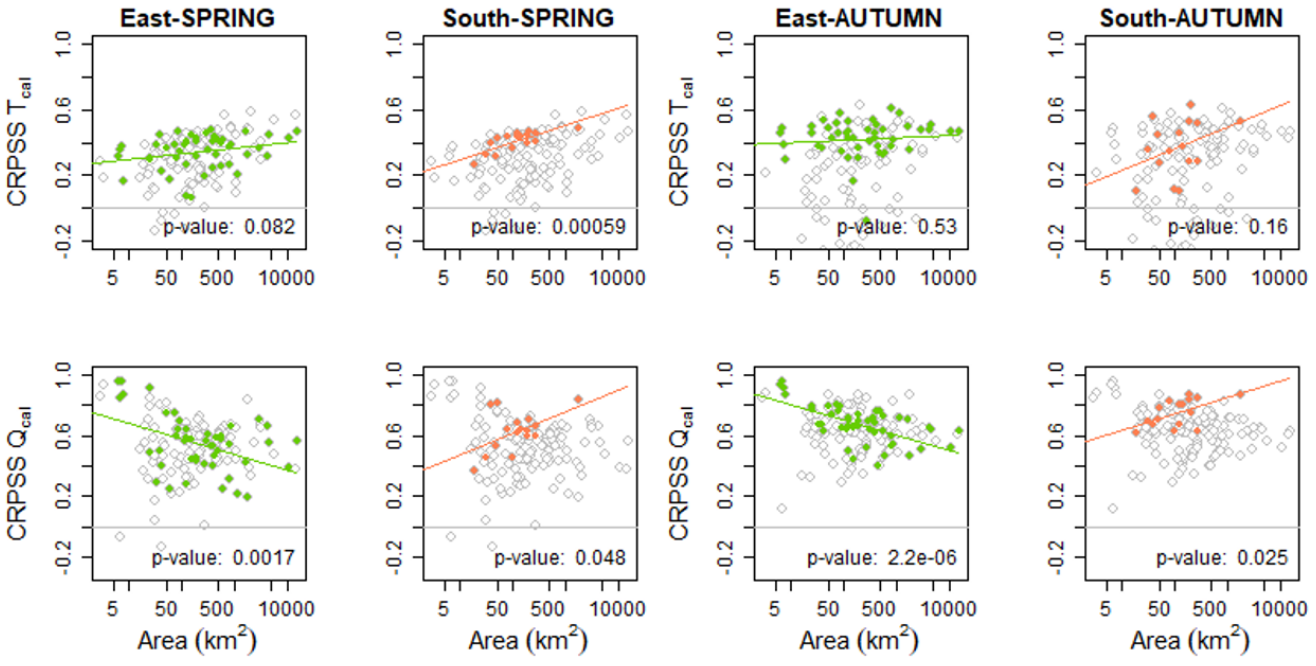

Figure 9: Temperature (top panels) and streamflow (bottom panels) CRPSS for the two regions East (E) and South (S), plotted as a function of catchments area for both autumn and spring. The colored dots show the CRPSS for the respective regions whereas the grey circles show the CRPSS for all 139 catchments. The linear regression line is plotted along with its p-value (significantly different from zero for p-values < 0.05). All results are presented for lead time 5 days.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-373
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 31 July 2018

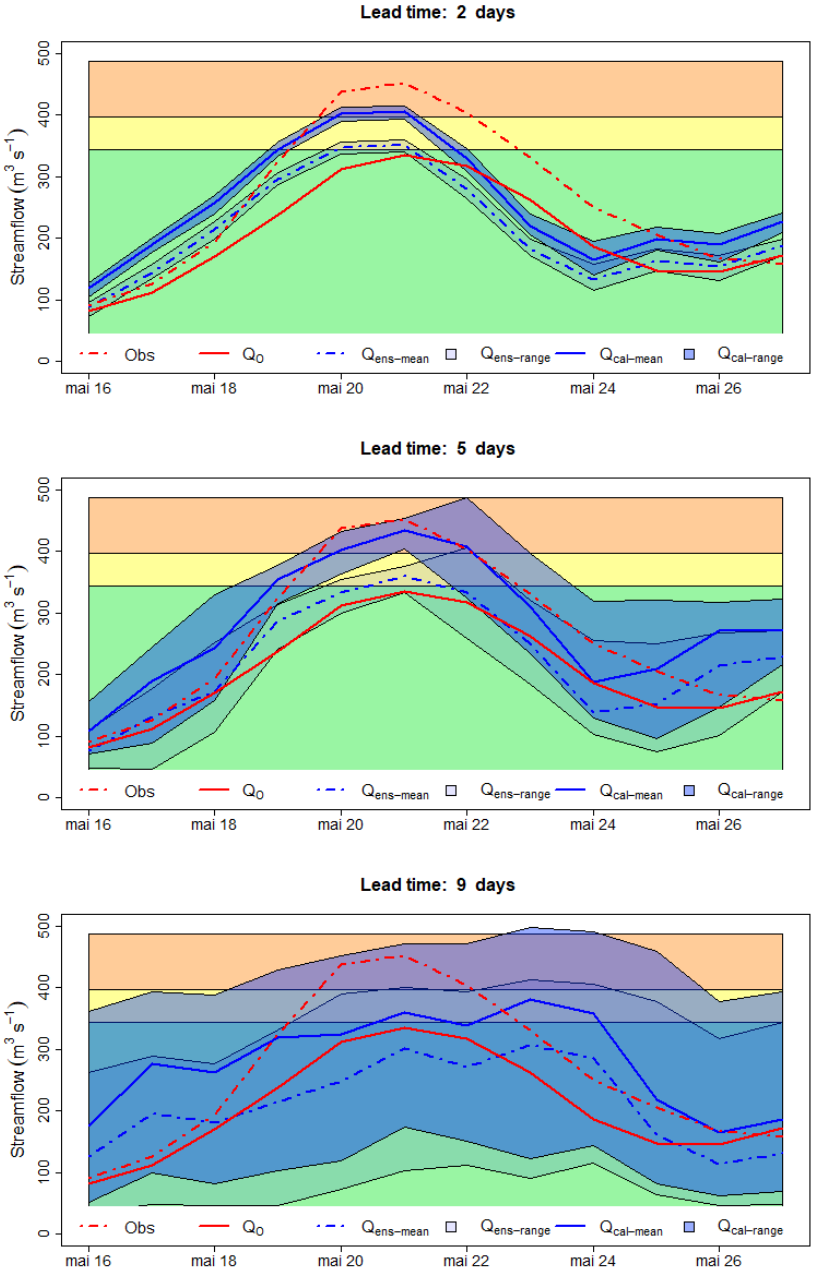

Figure 10: Forecasted streamflow for the Bulken catchment fort lead times 2, 5 and 9 days. Forecast issue dates on the x-axis, and streamflow (m^3s^{-1}) on y-axis. Observed streamflow (red dotted line), model streamflow with SeNorge observations (red solid line), mean Q_{cal} (blue line), ensemble range Q_{cal} (light blue area), ensemble mean Q_{ens} (blue dotted line) and ensemble range Q_{ens} (light grey area). The background colors indicate the warning level green, yellow, and orange.