Hydrology Research

© 2023 The Authors

Hydrology Research Vol 54 No 2, 116 doi: 10.2166/nh.2023.024

Pre- and postprocessing flood forecasts using Bayesian model averaging

Trine Jahr Hegdahl[a,*], Kolbjørn Engeland[a,b], Ingelin Steinsland[c] and Andrew Singleton[d]

[a] Norwegian Water Resources and Energy Directorate, Hydrological Modelling, Oslo 0301, Norway
[b] Department of Geosciences, University of Oslo, Oslo 0316, Norway
[c] Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim 7034, Norway
[d] Norwegian Meteorological Institute, Oslo 0313, Norway
*Corresponding author. E-mail: tjh@nve.no

TJH, 0000-0002-0940-0065

ABSTRACT

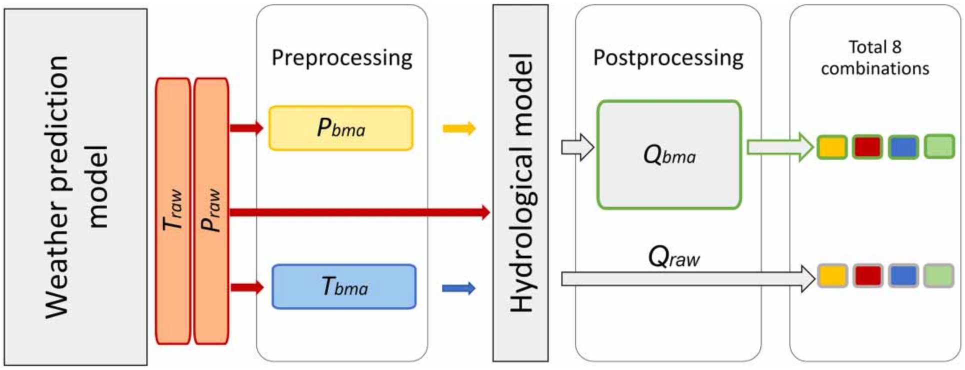
In this study, pre- and postprocessing of hydrological ensemble forecasts are evaluated with a special focus on floods for 119 Norwegian catchments. Two years of ECMWF ensemble forecasts of temperature and precipitation with a lead time of up to 9 days were used to force the operational hydrological HBV model to establish streamflow forecasts. A Bayesian model averaging processing approach was applied to preprocess temperature and precipitation forecasts and for postprocessing streamflow forecasts. Ensemble streamflow forecasts were generated for eight schemes based on combinations of raw, preprocessed, and postprocessed forecasts. Two datasets were used to evaluate the forecasts: (i) all streamflow forecasts and (ii) forecasts for flood events with streamflow above mean annual flood. Evaluations based on all streamflow data showed that postprocessing improved the forecasts only up to a lead time of 2–3 days, whereas preprocessing temperature and precipitation improved the forecasts for 50–90% of the catchments beyond 3 days' lead time. We found large differences in the ability to issue warnings between spring and autumn floods. Spring floods had predictability for up to 9 days for many events and catchments, whereas the ability to predict autumn floods beyond 3 days was marginal.

Key words: BMA, ensemble, flood, forecasting, postprocessing, preprocessing

HIGHLIGHTS

- The study evaluates the univariate and the combined effects of preprocessing both precipitation and temperature forecasts together with the postprocessing of streamflow.
- Evaluating forecasts of both floods as well as all streamflow values.
- Large catchment sample for more robust assessment of preferred processing approaches.
- Seasonal and regional differences in processing approaches are assessed.

GRAPHICAL ABSTRACT



INTRODUCTION

Early warnings based on flood forecasts enable both the management authorities and the public to take necessary measures to reduce the economical, personal, and social impact of floods (e.g., UNISDRI 2004; Pappenberger *et al.* 2017). However, in common with any sort of forecast, an inherent feature of flood forecasting is uncertainty. In the hydro-meteorological forecasting chain, the forecast uncertainty comes from multiple sources. There is uncertainty in observations, initial conditions, forcing data, model description, and model parameters (e.g., Buizza *et al.* 1999; Zappa *et al.* 2011).

To capture the uncertainty in weather prediction caused by initial conditions (e.g., Lorenz 1969) and model parametrization, ensemble prediction systems (EPS) were developed (e.g., Leith 1974; Buizza 2015). The use of hydrological ensemble forecasts has been studied in the literature, see, e.g., Cloke & Pappenberger (2009), Wetterhall *et al.* (2013). To get unbiased and reliable hydrological forecasts, preprocessing (applied to the meteorological forcing) and/or postprocessing (applied to the hydrological output) techniques are needed. For flood forecasting, important sources of uncertainty and errors are the precipitation and temperature forecasts (e.g., Zappa *et al.* 2011). These variables are considered for preprocessing in this paper.

For a national or regional flood forecasting service, a large number of catchments with different hydrological processes and regimes are considered. In most papers, ensemble forecasts of all streamflow values for one or a small number of catchments are evaluated. Therefore, to assess the added value of pre- and postprocessing on flood forecasts, a case study from a large number of catchments that well represent the variability of hydrological processes is needed to provide robust conclusions.

The quality of ensemble forecasts is often measured by the key characteristics' reliability and accuracy. A forecast is reliable (statistically calibrated) when, e.g., for 90% of the forecasts, the observations are within the 90% prediction interval. Raw forecast ensembles are often biased and underdispersive (Gneiting *et al.* 2005). A lack of dispersion in global meteorological ensembles is most evident for the shortest lead times and can be explained by slower growth rates of the perturbations in the ensemble prediction system compared to those of an instable 'true' atmosphere (Hamill 2001). To correct for bias and underdispersion in ensemble systems, different statistical postprocessing approaches are proposed, see Li *et al.* (2017) and Vannitsem *et al.* (2018) for comprehensive reviews. These approaches include both parametric approaches relying on parametric probability distributions, for example, Bayesian model averaging (BMA) and nonhomogeneous Gaussian regression (NGR), and nonparametric approaches like quantile regression and ensemble error dressing methods. In this study, we used BMA since it is well established and adapts easily to any kind of seasonality. Raftery *et al.* (2005) introduced BMA to the atmospheric community as a statistical method to achieve calibrated and sharp forecasts, and the method has since been widely used within the community (e.g., Fraley *et al.* 2010; Madadgar *et al.* 2014; Xu *et al.* 2019).

The effects of both pre- and postprocessing on short- to medium-range streamflow forecasts have been analyzed in previous studies (e.g., Zalachori *et al.* 2012; Roulin & Vannitsem 2015; Benninga *et al.* 2017; Sharma *et al.* 2018). Some key findings are that (i) calibrated precipitation forecasts do not necessarily lead to calibrated streamflow forecasts (Zalachori *et al.* 2012; Verkade *et al.* 2013; Benninga *et al.* 2017); (ii) postprocessing alone is the simplest way to improve forecasting performance (Zalachori *et al.* 2012; Sharma *et al.* 2018), but not always with a significant improvement (Benninga *et al.* 2017); (iii) preprocessing the meteorological forcing is important for forecasting high streamflows since errors from the meteorological model are dominant in this case (Benninga *et al.* 2017); (iv) preprocessing has the highest skill improvement in the warm season, whereas postprocessing is the most effective in the cold season with snow cover (Sharma *et al.* 2018). These findings indicate that the relative importance of pre- and postprocessing depends on factors including lead time, streamflow magnitude, and season. None of these studies have compared the univariate and the combined effects of including both precipitation and temperature forecasts in the preprocessing together with the postprocessing of streamflow on flood forecasts. Furthermore, these studies indicate that the effects depend on both climatological and physiographic catchment characteristics and that it can be useful to systematically evaluate the combination of pre- and postprocessing methods for a large set of catchments with variations of climatic and physiographic properties. In this study, we will evaluate (i) the univariate and the combined effects of preprocessing both precipitation and temperature forecasts together with the postprocessing of streamflow for forecasting floods as well as all streamflow values, and to (ii) perform the evaluation for a large catchment sample.

The main objective of this study is to assess the potential improvements in flood forecasts by combining pre- and postprocessing for a variety of catchments. Different schemes of pre- and postprocessing using BMA are evaluated within the operational flood forecasting setup used by the Norwegian flood forecasting service. The different schemes were tested for

119 catchments that vary in climatology, catchment characteristics, and hydrological regimes. During the study period, there were flood events in 80 of the catchments. The large number of flood events and catchments allowed us to provide robust assessments of the performance of the different schemes under different flood conditions.

The working hypothesis of this paper is that pre- and/or postprocessing improves streamflow forecasts and that the improvements differ between catchments and between events. We addressed the following questions:

1. How should pre- and postprocessing be combined to improve streamflow forecasts with an emphasis on floods?
2. Are there regional or seasonal patterns in the preferred combination of pre- and postprocessing?

STUDY AREA, HYDROLOGICAL MODEL, AND DATA

Area

Norway consists of several different climatic zones. The west coast of Norway forms a topographical barrier for the westerlies and orographic enhancement of precipitation makes this area one of the wettest parts of Europe, with an annual precipitation of around 4,000 mm. The driest regions have annual precipitation of around 400 mm (Hanssen-Bauer *et al.* 2017). The temperature depends on latitude, altitude, and distance from the coast.

The seasonal variation in runoff depends on seasonal variations in both temperature and precipitation. There are two basic runoff regimes in Norway. For coastal regions with a temperate climate, the highest flows occur during autumn and winter due to heavy rainfall. For inland regions with a sub-arctic or arctic climate, prolonged periods of winter temperatures below 0 °C result in a seasonal snow storage, winter low flow, and high streamflow during spring due to snowmelt. There are, however, many possible transitions between these two basic patterns (e.g., Gottschalk *et al.* 1979). We grouped the Norwegian catchments into five hydroclimatic regions (Figure 1) according to Hanssen-Bauer *et al.* (2017) and Vormoor *et al.* (2016).

The study area consists of 119 catchments distributed all over Norway (Figure 1). All selected catchments are part of the operational flood forecasting system and are mostly unregulated, with a large variation in size (3–15,447 km^2) and elevation (103–2,284 meters above sea level [m.a.s.l.]). Three catchments (Table 1) are presented in more detail to illustrate streamflow ensemble forecasts estimated by different processing approaches for three different flood events. The catchments were selected to represent the main flood-generating processes for the different regions. The catchments are all well described by the model.

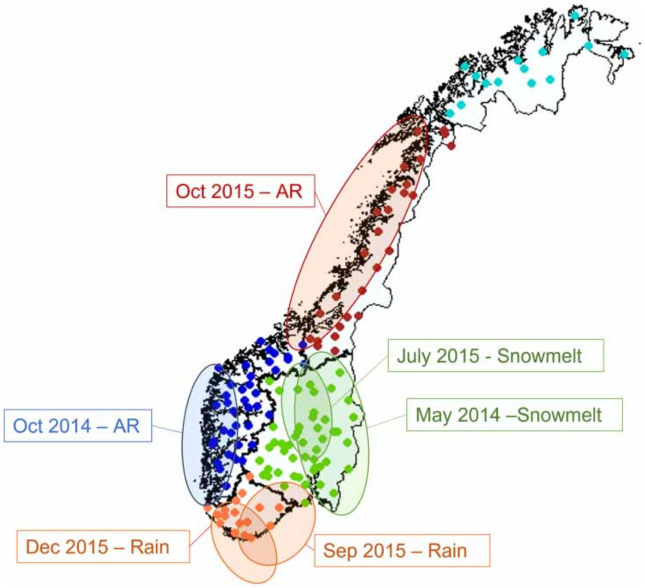


Figure 1 | The map shows the location of the outlet of the 119 catchments used in this study as well as a schematic overview of the areas affected by floods caused by different events (rain, snowmelt, and atmospheric river (AR)) during the study period 2014–2015. It is worth noting that not all catchments experienced floods within the areas. The colored dots indicate catchments by the regions, east (green), south (orange), west (blue), mid (dark red), north (light blue).

Table 1 | Catchment characteristics for selected catchments: catchment area, annual runoff (Q), catchment mean elevation (Mean elev), effective lake area (Eff lake), glacier area (Glacier)

Name	Area (km^2)	Annual *Q* (mm)	Mean elev (m.a.s.l)	Eff lake (%)	Glacier (%)	Selected Flood
Moeska	121	1,585	325	1.71	0.00	Rain: Dec 2015
Nybergsund	4,425	487	781	2.48	0.00	Snowmelt: May 2014
Bulken	1,092	2,038	867	0.88	0.39	AR: oct 2014

Hydrological model

We used the Hydrologiska Byråens Vattenbalance (HBV) model (Bergström 1976; Sælthun 1996; Beldring 2008) that is used by the operational flood forecasting service at the Norwegian Water Resources and Energy Directorate (NVE). The HBV model is a conceptual model whose vertical structure includes a snow routine, a soil moisture routine, and a response function that consists of two tanks. Quick runoff is represented by a nonlinear tank, whereas slow runoff is represented by a linear tank. The model divides each catchment into 10 elevation zones where each represents 10% of the catchment area. Catchment average temperature and precipitation are elevation adjusted using a catchment-specific lapse rate to attain one representative precipitation and temperature value for each elevation zone. The Nash–Sutcliffe efficiency (Nash & Sutcliffe 1970) and volume bias are used as calibration metrics. The calibration period, 1996–2012, gives a mean Nash–Sutcliffe 0.77 for all 119 catchments, with zero volume bias. The validation period, 1980–1995, shows a mean Nash–Sutcliffe of 0.73, with a mean volume bias of 5% (Gusong 2016).

Data

Meteorological observation SeNorge v1.1

We used the gridded daily temperature and precipitation data from the SeNorge v 1.1 dataset, which covers all of Norway with a 1×1 km grid size. The interpolation of observations to the grid is based on measured values at approximately 400 meteorological stations for precipitation, and 240 stations for temperature. Residual kriging is applied for spatial interpolation of detrended temperature values (Tveito 2007; Mohr 2008). Temperature is detrended by adjusting station data to sea level using a standard temperature lapse rate of 0.65 °C/100 m. Triangulation is used for the spatial interpolation of precipitation (Tveito 2007; Mohr 2008). The precipitation is further elevation corrected, using a constant increase of 10% per 100 m beneath 1,000 m.a.s.l, and 5% per 100 m above 1,000 m.a.s.l. (Tveito *et al.* 2005).

Meteorological forecasts ECMWF ENS

The temperature and precipitation forecasts used in the hydrological simulations of this study were taken from the European Center of Medium-Range Weather Forecast (ECMWF) forecast ensembles (ENS). ENS provides an ensemble of 51 members and a forecasting period of 246 h. The ensemble members are generated by adding small perturbations to the forecast initial conditions. The perturbations represent the uncertainty in the observations. Further, the uncertainty associated with the model physics is represented by perturbing the physics tendencies that come from the parametrizations and each member is perturbed individually. This method is known as the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and improves the forecasts giving a much better spread-error relationship compared to initial condition perturbations alone. A detailed description of the ECMWF ENS system is provided in, e.g., Buizza *et al.* (1999) and Persson (2015). The grid resolution of the model forecasts used in this study is 0.25° (i.e., model cycles/versions 40r1, and 41r1 (ECMWF 2018)). The variables used for the hydrological modeling are the accumulated precipitation and the 2-m temperature aggregated to catchment daily (06:00–06:00) mean values.

Streamflow reference simulations

To calibrate the hydrological model the streamflow measurements from the NVE database (https://www.nve.no/hydrology/) were used as a reference. To evaluate the streamflow forecasts, we used simulated streamflow (reference streamflow) created by running the hydrological model with SeNorge temperature and precipitation as forcing. Using this approach, we isolated the effect of the uncertainty in the weather forecasts, and we could ignore uncertainties in observed meteorological inputs, initial conditions, hydrological model parametrizations and parameters as suggested in Verkade *et al.* (2013).

Study period

The study period 2014 and 2015 was chosen since several large floods affected rivers in most parts of Norway during this period (Figure 1). In May 2014, there were large snowmelt floods in the central and eastern parts of Norway. In October 2014, western Norway was hit by an atmospheric river (a narrow plume of high moisture content air transported from the tropical and extratropical latitudes towards the poles, see, e.g., Zhu & Newell 1998), which led to the flooding of multiple rivers. Atmospheric rivers are responsible for extreme precipitation events when the moist air masses are orographically lifted at topographical barriers like the west coast of Norway (e.g., Stohl *et al.* 2008). In July 2015, there were snowmelt floods in central eastern Norway, and in September 2015, an extratropical cyclone, *Petra*, caused floods in Southern Norway. In early October 2015, a cyclone, *Roar*, caused floods in Trøndelag and Nordland and in early December a cyclone, *Synne,* caused floods in several catchments in south-west Norway, some exceeding the 200-year return level. Floods did not occur in all catchments; hence, the number of catchments used in the flood evaluation analysis was reduced to 80, from the original 119 catchments available for evaluation.

PRE- AND POSTPROCESSING

Processing chain

The temperature and precipitation forecast data from ECMWF were prepared by aggregating the variables from hourly to a daily time step. Thereafter the horizontal resolution was changed using nearest neighbor interpolation to a 1 × 1 km grid, equal to the SeNorge grid. For the temperature forecasts, a standard elevation adjustment of 0.65 °C/100 m was applied to account for the elevation differences between the original and the seNorge grid. Finally, the temperature and precipitation forecasts were aggregated to average values for each catchment. The ECMWF forecasts from 2014 and 2015 were used as forcing for the hydrological model, enabling a retrospective evaluation of the daily streamflow forecasts for almost 2 years. The unprocessed daily ensemble forecasts for each catchment are referred to as $T_{raw,t,l,s,m}$ and $P_{raw›t,l,s,m}$ where t is the issue time, l is the lead time, s is the catchment and m is the ensemble member.

We used BMA to process all ensembles of temperature, precipitation and streamflow. We chose BMA since it is flexible enough to adjust for biases and spread in the raw forecasts, and is well established and easy to implement. Using only BMA allowed us to address the main objective of this paper in a consistent way. BMA was applied to the raw forecasts to produce the preprocessed ensemble, with the members referred to as $T_{bma,t,l,s,m}$ and $P_{bma,t,l,s,m}$, where t is the issue time, l is the lead time, s is the catchment and m is the ensemble member. For postprocessing streamflow, we used BMA to create $Q_{bma,t,l,s,m}$. The processing was applied to each issue date, t, lead time l, and catchment, s, independently for all combinations. To improve readability, t,l,s,m is suppressed in the remainder of this paper. We evaluated all combinations. The four combinations of temperature and precipitation (T_{bma} and P_{bma} together with T_{raw} and P_{raw}) were run through the hydrological model resulting in four preprocessed streamflow forecasts (Q_{raw}). Thereafter, postprocessing the preprocessed forecasts resulted in four streamflow forecasts (Q_{bma}), which could be compared to Q_{raw} to establish the effect of postprocessing. See Figure 2 for an overview of the complete processing chain. A more detailed presentation of each step in the processing chain follows.

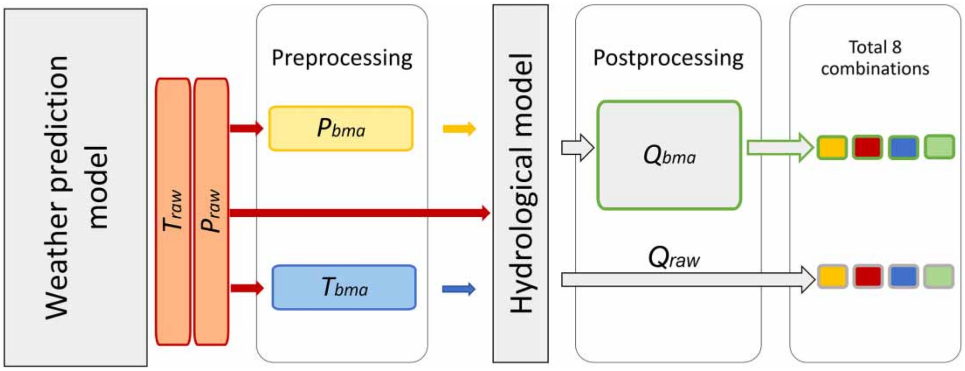


Figure 2 | The processing chain of the experimental set up. T_{raw} and P_{raw} are the unprocessed forecasts. The preprocessing producing the ensembles T_{bma} and P_{bma}. All combinations of T_{bma} and P_{bma} together with T_{raw} and P_{raw} were run through the hydrological model. BMA was further applied to the streamflow forecasts producing the ensembles Q_{bma} in addition to Q_{raw}. In total, eight combinations of pre- and postprocessing were evaluated. The PS were applied to each issue date, lead time, and catchment.

Bayesian model averaging

BMA aims to correct dispersion errors in a bias-corrected ensemble (Raftery *et al.* 2005). For each lead time, BMA uses a mixture distribution, where for an ensemble with M members, the density function conditioned on all ensemble members is the weighted average of kernels for each member m. The preprocessed meteorological ensembles were established by randomly drawing M realizations from the mixture distribution estimated by BMA. The kernel, for the quantity one wishes to forecast, y, is denoted by $f_\theta(y|x_m)$ where f is the kernel probability density function (pdf) with parameters θ, and x_m is the raw forecast's ensemble member. The pdf conditioned on all M ensemble members is the weighted average of the pdf for each member:

$$f(y|x_1, \ldots, x_M) \sim \sum_{m=1}^{M} w_m f_\theta(y|x_m) \tag{1}$$

where $\sum_{m=1}^{M} w_m = 1$ and the weights are interpreted as the posterior probabilities of each ensemble member. The ensembles in this paper are based on ECMWF ENS which comprises members that are considered exchangeable, and weights and parameters can be constrained to be equal for all members (Fraley *et al.* 2010). For each issue date, we used the previous n days of ensemble forecasts and reference observations to estimate the parameters in the kernel. To account for the specific properties of temperature, precipitation and streamflow, different kernel distributions were used, and the details are provided below.

BMA for temperature (T_{bma})

We followed Raftery *et al.* (2005) and used a Normal distribution as the kernel for the temperature BMA models. Since the temperature ensemble forecasts were not already bias corrected, the mean is specified as $a_0 + a_1 T_{raw,m}$, where $T_{raw,m}$ is the temperature forecast for ensemble member m and a_0 and a_1 are regression parameters that account for any bias. The parameters are specific for each catchment, issue date, and lead time and are the same for all ensemble members.

$$f\left(T_{bma}|T_{\mathrm{raw,m}}\right) \sim \mathrm{N}\left(a_0 + a_1 T_{raw,m}, \sigma^2\right), \tag{2}$$

To estimate the parameters a_0, a_1, and σ in Equation (2), the catchment average temperatures from SeNorge were used as a reference.

BMA for precipitation (P_{bma})

We followed Sloughter *et al.* (2007) who proposed a Bernoulli-gamma distribution as the kernel in the BMA precipitation models to establish P_{bma}.

$$f(P_{bma}|P_{raw,m}) = f(P_{bma} = 0|P_{raw,m}) I_{\{P_{bma}=0\}} + f(P_{bma} > 0|P_{raw,m}) h(P_{bma}|P_{raw,m}) I_{\{P_{bma}>0\}} \tag{3}$$

where $I_{\{\}}$ is unity if the condition within the brackets is true and zero otherwise. $f(P_{bma} = 0|P_{raw,m})$ is the probability of zero precipitation given by a logistic regression model:

$$f(P_{bma} = 0|P_{raw,m}) = \frac{1}{1 + exp(b_0 + b_1 P_{raw,m}^{1/3} + b_2 \delta_m)} \tag{4}$$

where b_0, b_1, and b_2 are regression parameters common for all ensemble members and δ_m equals 1 if $P_{raw,m} = 0$ and equals 0 otherwise.

$h(P_{bma}|P_{raw,m})$ was assumed to follow a gamma distribution for the cube root transformation $P'_{bma} = P_{bma}^{1/3}$ of the precipitation, where the mean (μ_m) and variance (σ_m^2) of the distribution depend on the ensemble member:

$$\mu_m = c_0 + c_1 P_{raw,m}^{1/3} \text{ and } \sigma_m^2 = d_0 + d_1 P_{raw,m} \tag{5}$$

where all parameters c_0 and c_1, d_0 and d_1 were the same for all ensemble members. The seven parameters in the Bernoulli-gamma kernels were estimated using the catchment average precipitation from seNorge as a reference.

BMA for streamflow (Q_{bma})

We applied a Box–Cox transformation (Box & Cox 1964; Duan *et al.* 2007) on both observed and forecasted streamflow to create the transformed streamflow Q' normally distributed:

$$Q' = \begin{cases} \frac{(Q^{\lambda} - 1)}{\lambda} & for\ \lambda \neq 0 \\ log(Q) & for\ \lambda = 0 \end{cases} \tag{6}$$

where λ is a transformation parameter. The Box–Cox transformation has proven valuable for hydrological applications (e.g., Bates & Campbell 2001; Thyer *et al.* 2002; Yang *et al.* 2007; Engeland *et al.* 2010). We used a fixed value for λ supported by previous studies by Engeland *et al.* (2010), who found that $\lambda = 0.2$ gave forecast errors that were approximately independent of forecasted values. As for temperature, we applied the BMA with a combination of normal kernels for postprocessing the streamflow forecasts, such that

$$f\left(Q'_{bma}|Q'_{\mathrm{raw,m}}\right) \sim \mathrm{N}\left(a_0 + a_1 Q'_{raw,m}, \sigma^2\right) \tag{7}$$

BMA training length

Following Raftery *et al.* (2005), the BMA models for temperature, precipitation, and streamflow were trained on data from a time window prior to the issue date for each forecast. We tested different training lengths for all variables and lead times, using CRPS (description in the following section) as the evaluation metric. Experiments with different training lengths showed that the optimal window size depends on variable, lead time, and whether CRPS was calculated for all data or only for days with flooding. Precipitation was most sensitive to the training length and a 45-day training period was found to be optimal for most catchments and lead times. To maintain consistency during the evaluation we used a 45-day training period for all variables (i.e., temperature, precipitation, and streamflow).

Temperature and precipitation dependence structure (ensemble copula coupling)

The BMA models described above were applied independently to each weather variable, each location (here catchment) and each lead time. The preprocessed ensembles were established by drawing 51 new realizations from the mixture distribution of each BMA model independently. To recreate forecast trajectories of temperature and precipitation, it is necessary to account for the temporal and inter-variable dependence structures. In this study, it was achieved by using an approach similar to Ensemble Copula Coupling (ECC, Schefzik *et al.* 2013). The original 51 ensemble members (o,m) for temperature and precipitation were, for each location, issue date, and lead time, assigned a rank ($r_{o,m}$), where o refers to the original ensemble member. Similarly, the 51 BMA-processed precipitation and temperature ensemble members were assigned a rank ($r_{n,m}$), where n,m refers to the BMA-processed ensemble member. The 51 preprocessed ensemble members were reordered by using $r_{o,m}$ and $r_{n,m}$ as keys to keep the preprocessed ensemble members in the same rank sequence as the original ensemble members. By applying this method to all variables, lead times, and issue dates we maintain the dependency between the variables, as well as the temporal dependency for each of the variables.

EVALUATION

We evaluated the pre- and postprocessing methods for the study period using both the full dataset and the flood dataset using continuous rank probability score (CRPS), skill score (CRPSS) and the critical success index (CSI) as evaluation metrics.

CRPS and continuous rank probability skill score (CRPSS)

The continuous rank probability score (CRPS) has properties that are appealing for the evaluation of an ensemble forecast. CRPS will give credit to high probabilities close to the reference, which is not necessarily the case for other ensemble verification scores (Gneiting *et al.* 2007). CRPS has the same unit as the observations (m^3/s for streamflow), and is negatively oriented, where zero is the optimal value. For a deterministic forecast, CRPS reduces to the mean absolute error (MAE, Hersbach 2000), which enables a comparison between a deterministic and an ensemble forecast. CRPS measures the integral of squared difference between the forecast and the observation, both given as a cumulative distribution function (cdf). If the observation is deterministic the Heaviside function is used for the observation cdf (Hersbach 2000). For ensemble forecasts, the CRPS is calculated discretely since both the observations and the forecasts are reported in discrete intervals (Hersbach 2000, Equation (8)):

$$\mathrm{CRPS} = \frac{1}{M}\sum_{m=1}^{M}|x_m - x_{obs}| - \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}|x_m - x_n| \tag{8}$$

where M is the ensemble size, x_m is ensemble member m and x_{obs} is the reference observation. For a time-series of forecasts, the mean CRPS for each scheme ($\overline{\mathrm{CRPS}_{\mathrm{PS}}}$) can be calculated.

The continuous ranked probability skill score (CRPSS, Equation (9)) enables assessment of the skill of the different processing schemes (PS) relative to the raw forecasts (raw). The mean CRPS for each scheme ($\overline{\mathrm{CRPS}_{\mathrm{PS}}}$) and for the unprocessed forecasts ($\overline{\mathrm{CRPS}_{\mathrm{raw}}}$) are used to calculate CRPSS.

$$\mathrm{CRPSS}_{\mathrm{PS}} = 1 - \frac{\overline{\mathrm{CRPS}_{\mathrm{PS}}}}{\overline{\mathrm{CRPS}_{\mathrm{raw}}}} \tag{9}$$

Note that CRPSS has 1 as the optimal value and is positively oriented. Since CRPSS has no units, we could calculate average skill scores across all catchments. CRPS and CRPSS were calculated for the complete dataset as well as for the flood dataset.

Critical success index

In an operational flood forecasting setting, flood warnings are issued when there is a certain probability for streamflow to exceed predefined flood warnings thresholds. The occurrence and nonoccurrence of floods are therefore binary events that can be summarized in a contingency table (Table 2) providing an overview of hits (H), missed events (M), false alarms (F), and correct nonevents (N). Based on the contingency table shown in Table 3, the following indices can be used to evaluate the performance of a forecasting system.

Hit ratio, where a hit rate of 1 is the best performance (S_{R}): $S_{\mathrm{R}} = \dfrac{H}{H+M}$ (10)

False alarm ratio (F_{R}): $F_{\mathrm{R}} = \dfrac{F}{H+F}$ (11)

Critical Success Index (CSI): $\mathrm{CSI} = \dfrac{H}{H+F+M}$ (12)

Since floods are rare events, there are a small number of flood events compared to the number of nonevents. A good forecast has a high hit ratio and a low false alarm ratio. The CSI (Jolliffe & Stephenson 2012) balances these two aims by penalizing the hit ratio for both the missed events (M) and the false alarms (F). The CSI has a value between zero and one, with one being the optimal value. In an operational setting, a warning will be issued when a predefined number of ensemble members (or a defined probability) exceeds the flood warning threshold. For the simplicity of this work, we have chosen a limit of 10 members exceeding the mean annual flood level. The mean annual flood has a return period of 2.33 years (i.e., ~20% probability of occurrence).

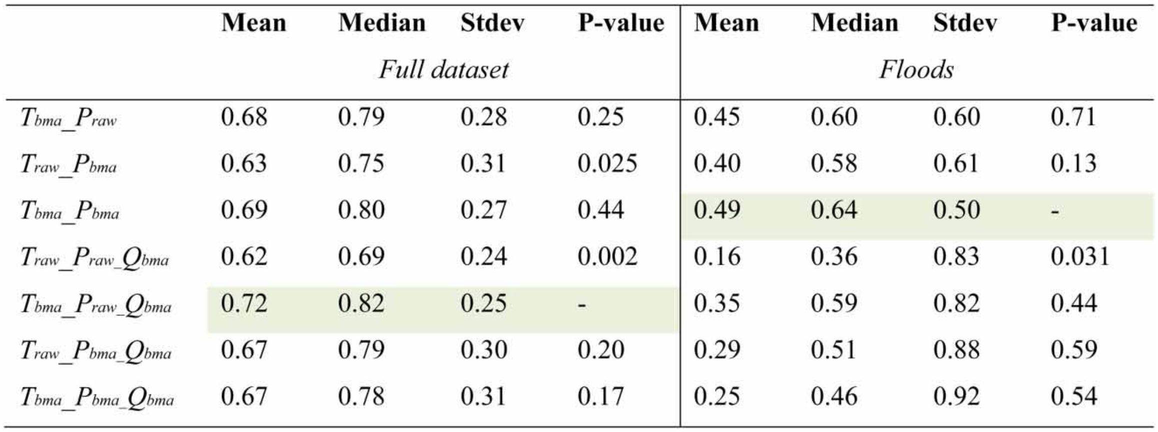
Floods by seasons

The performance of flood forecasts can differ between seasons for several reasons. One reason is that flood-dominating processes often are aligned to season, e.g., snowmelt contribution to floods dominates in spring, and rain-induced floods

Table 2 | Contingency table for classification of hits (H), missed events (M), false alarms (F), and correct nonevents (N)

		Observation	
		No	**Yes**
Forecast	**No**	N	M
	Yes	F	H

Table 3 | Summary statistics of CRPSS values for a lead time of 5 days for the full dataset and the flood dataset



	Mean	Median	Stdev	P-value	Mean	Median	Stdev	P-value
	Full dataset				*Floods*			
$T_{bma}\_P_{raw}$	0.68	0.79	0.28	0.25	0.45	0.60	0.60	0.71
$T_{raw}\_P_{bma}$	0.63	0.75	0.31	0.025	0.40	0.58	0.61	0.13
$T_{bma}\_P_{bma}$	0.69	0.80	0.27	0.44	0.49	0.64	0.50	-
$T_{raw}\_P_{raw}\_Q_{bma}$	0.62	0.69	0.24	0.002	0.16	0.36	0.83	0.031
$T_{bma}\_P_{raw}\_Q_{bma}$	0.72	0.82	0.25	-	0.35	0.59	0.82	0.44
$T_{raw}\_P_{bma}\_Q_{bma}$	0.67	0.79	0.30	0.20	0.29	0.51	0.88	0.59
$T_{bma}\_P_{bma}\_Q_{bma}$	0.67	0.78	0.31	0.17	0.25	0.46	0.92	0.54

The *p*-values show the outcome of a *t*-test comparing each processing approach to the best ones. For the full dataset, $T_{bma}\_P_{raw}\_Q_{bma}$ is the best whereas for the flood dataset, $T_{bma}\_P_{bma}$ is the best.

dominate in autumn. Another example are seasonal dependent biases, for example, a negative bias in the temperature ensemble forecast in autumn and winter for the Norwegian west coast (Seierstad *et al.* 2016; Hegdahl *et al.* 2019). For these reasons, we divided the flood events into spring and autumn floods and used CSI to evaluate how the performance of processing methods depends on the season. We defined spring from April 4 to June 13, and autumn from September 1 to December 10.

RESULTS

Skill – relations to lead time for all data and floods

We used CRPSS, with the raw ensembles as a benchmark, to evaluate how the different processing approaches affected the performance of ensemble streamflow forecasts for all lead times and catchments, for the full data set and for the subset of floods. CRPSS for all data and catchments (Figure 3 left and Table 3) show that nearly all catchments have a CRPSS above zero and therefore benefit from processing and that postprocessing in combination with preprocessing is most important for the short lead times, whereas postprocessing alone gives the lowest CRPSS. Preprocessing of temperature alone or combined with preprocessing of precipitation are the two best approaches for a lead time of 9 days. The *t*-test in Table 3 shows that it is difficult to find one method that is significantly better than all the others for all of Norway. The best processing approach ($T_{bma}\_P_{raw}\_Q_{bma}$) is significantly better than preprocessing only precipitation $T_{raw}\_P_{bma}$ or postprocessing streamflow without any preprocessing ($T_{raw}\_P_{raw}\_Q_{bma}$).

The variability in CRPSS is larger for the flood dataset (Figure 3 right and Table 3) compared to the full dataset, meaning that the benefit from the PS under flood conditions is not so high for all catchments, and for several catchments, the forecasts worsen (those where CRPSS is below zero). For the flood dataset, we find that if only preprocessing is applied, preprocessing both precipitation and temperature gives the highest skill. For the approaches including postprocessing, we see that postprocessing alone is the worst processing scheme, and that combining preprocessing of temperature with postprocessing is the best approach for more catchments. For the longer lead times, there are increasingly more catchments where postprocessing leads to a poorer performance, compared to using the raw forecast (our reference forecast). The *t*-test in Table 3 shows that it is difficult to find one method that is significantly better than all the others for all of Norway. The best processing approaches for the flood data ($T_{bma}\_P_{bma}$) and for all data ($T_{bma}\_P_{raw}\_Q_{bma}$) are both significantly better than postprocessing streamflow without any preprocessing ($T_{raw}\_P_{raw}\_Q_{bma}$).

CRPS – relations to location for the flood dataset

In Figure 4, maps of the processing approaches that achieve the best performance according to CRPS for the flood dataset are shown for lead times of 1, 5, and 9 days. The same results are summarized in Table 4 including summaries for each

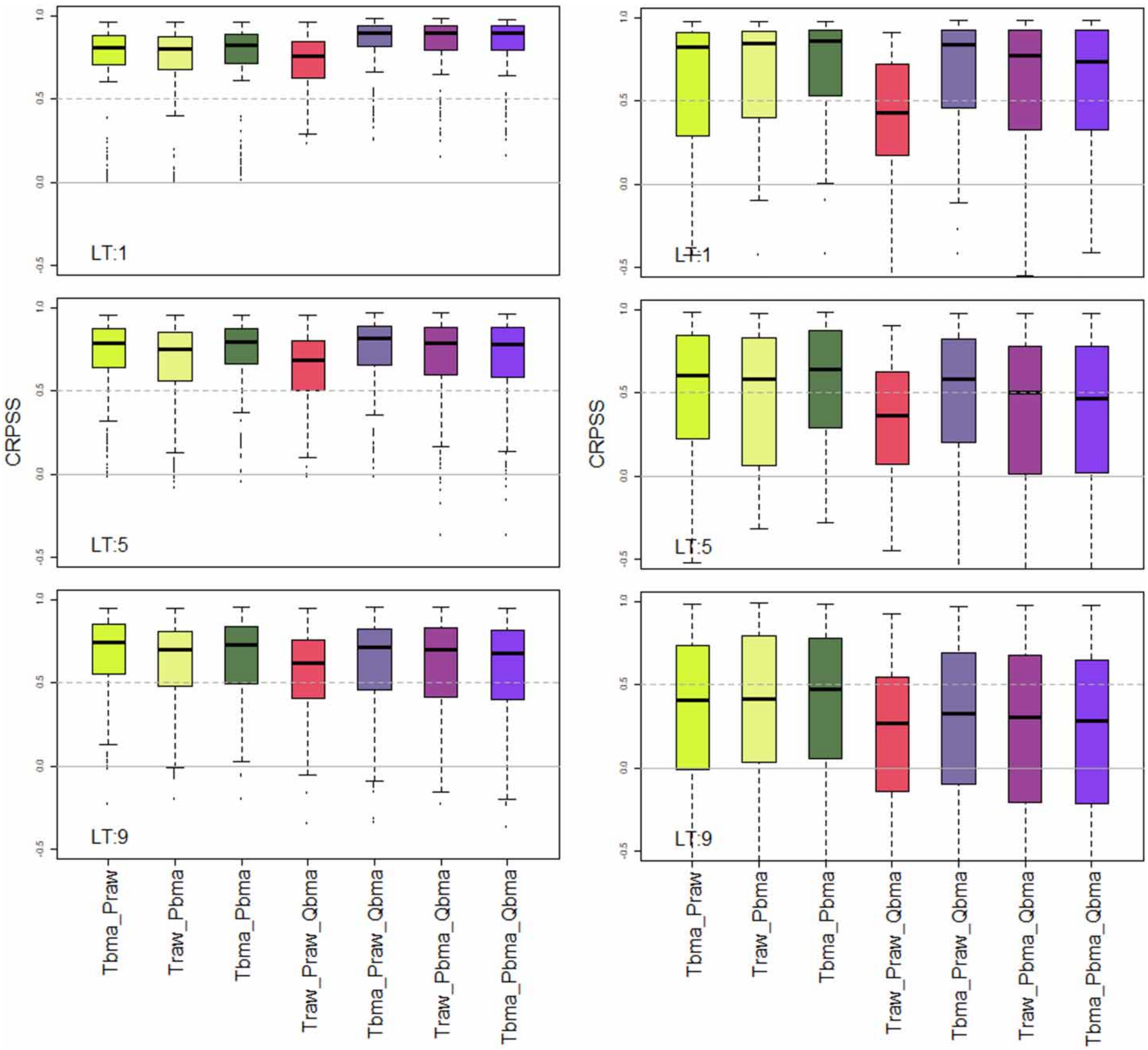


Figure 3 | Boxplot of CRPSS (optimal value is 1) for all catchments based on the full dataset (left) and the flood event dataset (right) for all PS (*x*-axis) and lead times of 1, 5, and 9 days (rows). The first three boxplots indicate the different preprocessing schemes, whereas the last four indicate PS that include a postprocessing step.

hydroclimatic region identified in Figure 1 (east, south, west, mid, and north). In the left column, we show which of the preprocessing approaches resulted in the best CRPS. We see that P_{bma}, alone or together with T_{bma}, gives the best results for western and southern coasts of Norway for lead times of 1 and 5 days, whereas for 9 days lead time, T_{bma} alone is more important. The success of the P_{bma}, in the coastal regions, could be that the floods are mainly rain driven. T_{bma} has a less clear spatial pattern. The benefit of processing decreases with lead time as the number of catchments with the best performance for the raw forecasts increases with lead time. In the middle column of Figure 4, we show the best postprocessing schemes for catchments where including postprocessing gave the best performance. Here we see that Q_{bma} alone is the most successful for lead time of 1 day, whereas the combination of T_{bma} and Q_{bma} dominates for 5 days. The T_{bma} and Q_{bma} are the least successful processing in eastern Norway. The right column in Figure 4 shows if a scheme including only preprocessing or both pre- and postprocessing performed the best. We see that a majority of catchments located inland, at high elevations or in eastern Norway benefit from postprocessing for lead times of 1 and 5 days, whereas the coastal catchments benefit to a smaller degree from postprocessing. The benefit of processing decreases with lead time.

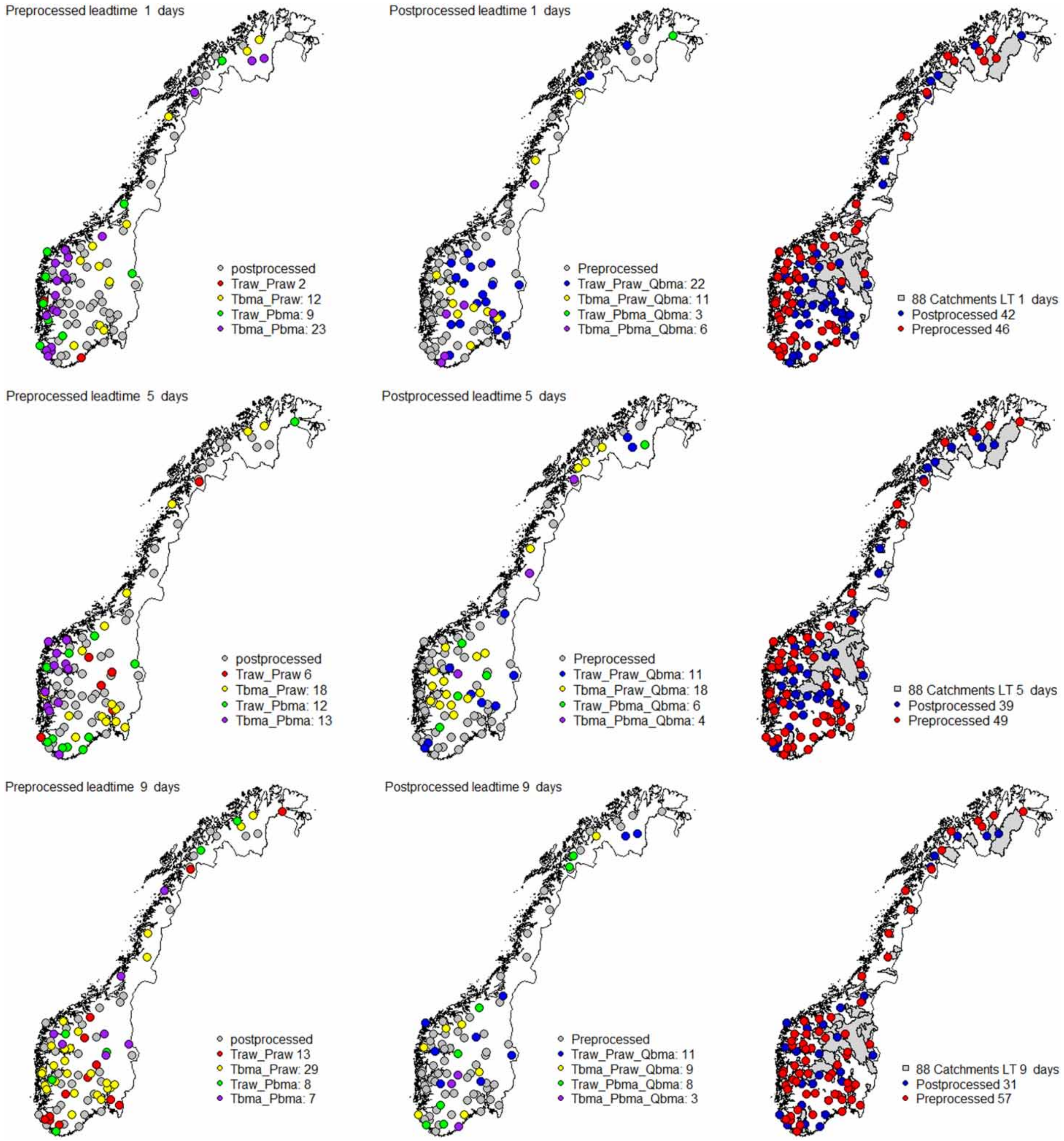


Figure 4 | The maps in the left column show the catchments where the different preprocessing schemes provide the best flood forecast. The middle column shows postprocessing schemes that provide the best CRPS. Figures to the right indicate catchments where any preprocessing approaches alone (red dots) or the combination of pre- and postprocessing (blue dots) provides the highest performance. All evaluation of CRPS was applied for the subset of floods, and by the mean CRPS for lead times of 1, 5, and 9 days.

Although Q_{bma} alone is the best approach for lead time of 1 day in a large proportion of the catchments (22 of 88), in particular in eastern Norway (8 of 26) (Table 4), it has the worst average performance since it results in low, and even negative CRPSS values in several catchments (Figure 3, right column). This indicates that Q_{bma} alone lacks robustness.

Table 4 | For each processing scheme, the number of catchments giving the best CRPS are presented for lead times of 1, 5, and 9 days and sorted by the hydroclimatic regions East (E), South (S), West (W), Mid (M), and North (N), see Figure 1

Lead time	*1 day*					∑	*5 days*					∑	*9 days*					∑
Region	*E*	*S*	*W*	*M*	*N*		*E*	*S*	*W*	*M*	*N*		*E*	*S*	*W*	*M*	*N*	
Traw_Praw	1	1	0	0	0	2	2	1	3	0	0	6	5	3	2	2	1	13
Tbma_Praw	4	3	5	0	0	12	6	1	6	2	3	18	11	3	13	2	1	30
Traw_Pbma	2	2	4	0	1	9	4	4	3	0	1	12	1	1	2	1	3	8
Tbma_Pbma	5	3	12	2	0	22	3	2	8	0	0	13	2	1	4	0	0	7
∑preprocessed	12	9	21	2	1	45	15	8	20	2	4	49	19	8	21	5	5	58
Traw_Praw_Qbma	8	1	3	4	6	22	2	4	4	0	1	11	1	3	6	0	1	11
Tbma_Praw_Qbma	4	2	4	0	1	11	4	2	5	4	3	18	2	2	2	1	2	9
Traw_Pbma_Qbma	1	1	2	0	0	4	3	1	2	0	1	7	3	1	2	1	0	7
Tbma_Pbma_Qbma	1	2	1	1	1	6	2	0	0	1	0	3	1	1	0	0	1	3
∑postprocessed	14	6	10	5	8	43	11	7	11	5	5	39	7	7	10	2	4	30

CSI for the whole year, spring, and autumn floods

In this evaluation, the processing scheme giving the highest critical success index (CSI) for each catchment is considered, and the number of catchments for which the specific scheme gave the best CSI is in Figure 5. The CSI value indicates the skill of the forecasts. No hits will give a CSI of zero, whereas all hits and no missed events or false alarms will give a CSI of one. For each catchment, multiple methods can achieve equal CSI and the number of 'best' CSI can exceed the total number of catchments.

Evaluating CSI for floods from the whole year did not give any clear indication as to which of the processing methods was better at predicting floods. This might be caused by floods being generated from rain, snowmelt, or a combination of those. However, by separating the flood dataset between floods occurring in spring and those occurring in autumn (Figure 5) we

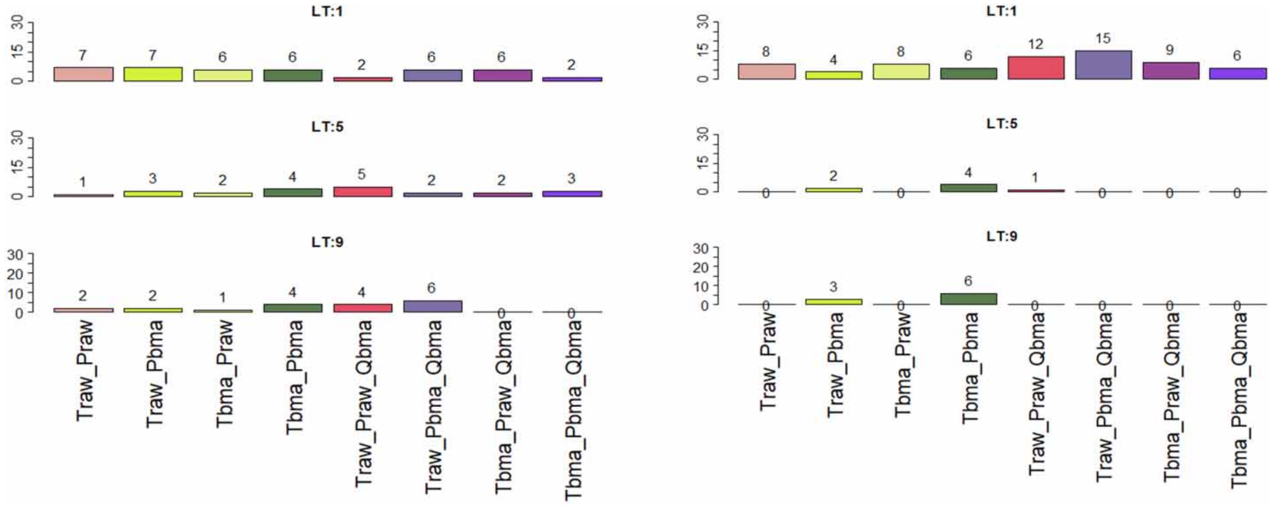
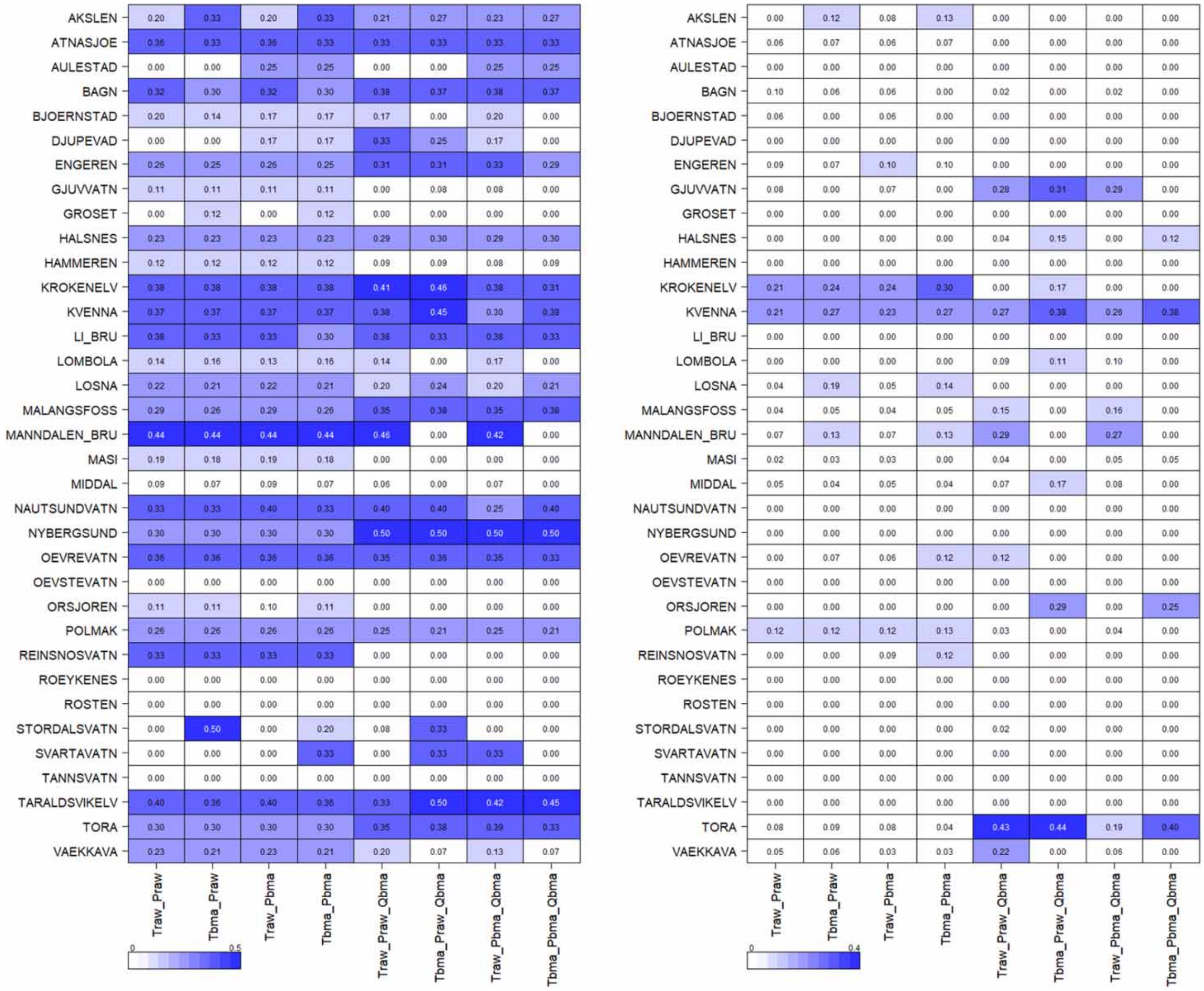


Figure 5 | Number of catchments (vertical axis) for which the schemes on the horizontal axis gave the best critical success index (CSI) for spring (left) and autumn (right) floods. Each row represents one lead time (1, 5, and 9 days) and includes all PS. A value of zero indicates that this method was not the best method for any of the catchments. If all schemes resulted in CSI of zero for one catchment, this catchment was not counted.

attain some interesting insight. For spring (Figure 5, left) we see that for a lead time of 1 day, the number catchments for which the different processing method performed the best is almost similar, indicating several successful methods. Q_{bma} alone or in combination with T_{bma} and Q_{bma} were the least successful methods. For lead times of 5 and 9 days, we see some improvement by applying pre- and/or postprocessing to spring floods.

For autumn (Figure 5, right) the results differ from the spring results. For a lead time of 1 day, the predictions are improved in several catchments by including postprocessing. Postprocessing has zero predictability (CSI is zero) for most of the catchments for lead times of 5 and 9 days. Only a few catchments have better predictive skill when applying P_{bma} alone or in combination with T_{bma}.

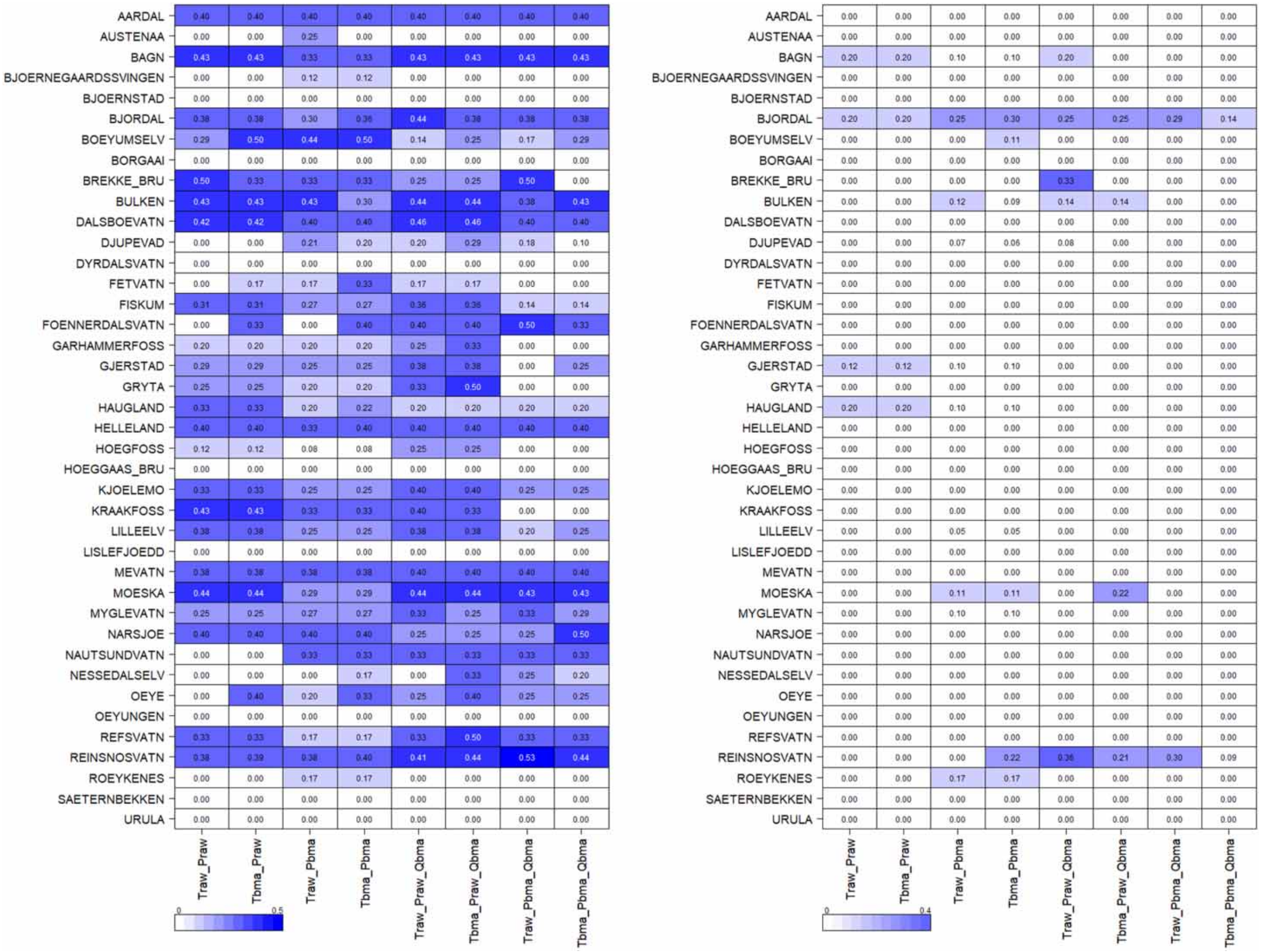
In Figures 6 and 7, the CSI values for each catchment and all PS are presented for spring and autumn floods respectively. For spring, lead times of 1 day (left) and 9 days (right) are presented in Figure 6, and for autumn lead times of 1 day (left) and 3 days (right) are presented in Figure 7. A white space indicates that for the actual catchment and processing scheme the floods were not forecasted, i.e., zero hits. We see that the highest CSI is 0.50, whereas for several cases, none of the processing approaches resulted in hits. In particular, for autumn floods at a lead time of 3 days, there are 33 catchments where the raw forecasts gave no hits, none of the processing approaches helped for 28 of these catchments.



	Traw_Praw	Tbma_Praw	Traw_Pbma	Tbma_Pbma	Traw_Praw_Qbma	Tbma_Praw_Qbma	Traw_Pbma_Qbma	Tbma_Pbma_Qbma
AKSLEN	0.20	0.33	0.20	0.33	0.21	0.27	0.23	0.27
ATNASJOE	0.36	0.33	0.36	0.33	0.33	0.33	0.33	0.33
AULESTAD	0.00	0.00	0.25	0.25	0.00	0.00	0.25	0.25
BAGN	0.32	0.30	0.32	0.30	0.38	0.37	0.38	0.37
BJOERNSTAD	0.20	0.14	0.17	0.17	0.17	0.00	0.20	0.00
DJUPEVAD	0.00	0.00	0.17	0.17	0.33	0.25	0.17	0.00
ENGEREN	0.26	0.25	0.26	0.25	0.31	0.31	0.33	0.29
GJUVVATN	0.11	0.11	0.11	0.11	0.00	0.08	0.08	0.00
GROSET	0.00	0.12	0.00	0.12	0.00	0.00	0.00	0.00
HALSNES	0.23	0.23	0.23	0.23	0.29	0.30	0.29	0.30
HAMMEREN	0.12	0.12	0.12	0.12	0.09	0.09	0.08	0.09
KROKENELV	0.38	0.38	0.38	0.38	0.41	0.46	0.38	0.31
KVENNA	0.37	0.37	0.37	0.37	0.38	0.45	0.30	0.39
LI_BRU	0.38	0.33	0.33	0.30	0.38	0.33	0.38	0.33
LOMBOLA	0.14	0.16	0.13	0.16	0.14	0.00	0.17	0.00
LOSNA	0.22	0.21	0.22	0.21	0.20	0.24	0.20	0.21
MALANGSFOSS	0.29	0.26	0.29	0.26	0.35	0.38	0.35	0.38
MANNDALEN_BRU	0.44	0.44	0.44	0.44	0.46	0.00	0.42	0.00
MASI	0.19	0.18	0.19	0.18	0.00	0.00	0.00	0.00
MIDDAL	0.09	0.07	0.09	0.07	0.06	0.00	0.07	0.00
NAUTSUNDVATN	0.33	0.33	0.40	0.33	0.40	0.40	0.25	0.40
NYBERGSUND	0.30	0.30	0.30	0.30	0.50	0.50	0.50	0.50
OEVREVATN	0.36	0.36	0.36	0.36	0.35	0.36	0.35	0.33
OEVSTEVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ORSJOREN	0.11	0.11	0.10	0.11	0.00	0.00	0.00	0.00
POLMAK	0.26	0.26	0.26	0.26	0.25	0.21	0.25	0.21
REINSNOSVATN	0.33	0.33	0.33	0.33	0.00	0.00	0.00	0.00
ROEYKENES	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROSTEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
STORDALSVATN	0.00	0.50	0.00	0.20	0.08	0.33	0.00	0.00
SVARTAVATN	0.00	0.00	0.00	0.33	0.00	0.33	0.33	0.00
TANNSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TARALDSVIKELV	0.40	0.36	0.40	0.36	0.33	0.50	0.42	0.45
TORA	0.30	0.30	0.30	0.30	0.35	0.38	0.39	0.33
VAEKKAVA	0.23	0.21	0.23	0.21	0.20	0.07	0.13	0.07

	Traw_Praw	Tbma_Praw	Traw_Pbma	Tbma_Pbma	Traw_Praw_Qbma	Tbma_Praw_Qbma	Traw_Pbma_Qbma	Tbma_Pbma_Qbma
AKSLEN	0.00	0.12	0.08	0.13	0.00	0.00	0.00	0.00
ATNASJOE	0.06	0.07	0.06	0.07	0.00	0.00	0.00	0.00
AULESTAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BAGN	0.10	0.06	0.06	0.00	0.02	0.00	0.02	0.00
BJOERNSTAD	0.06	0.00	0.06	0.00	0.00	0.00	0.00	0.00
DJUPEVAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ENGEREN	0.09	0.07	0.10	0.10	0.00	0.00	0.00	0.00
GJUVVATN	0.08	0.00	0.07	0.00	0.28	0.31	0.29	0.00
GROSET	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HALSNES	0.00	0.00	0.00	0.00	0.04	0.15	0.00	0.12
HAMMEREN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KROKENELV	0.21	0.24	0.24	0.30	0.00	0.17	0.00	0.00
KVENNA	0.21	0.27	0.23	0.27	0.27	0.38	0.26	0.38
LI_BRU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LOMBOLA	0.00	0.00	0.00	0.00	0.09	0.11	0.10	0.00
LOSNA	0.04	0.19	0.05	0.14	0.00	0.00	0.00	0.00
MALANGSFOSS	0.04	0.05	0.04	0.05	0.15	0.00	0.16	0.00
MANNDALEN_BRU	0.07	0.13	0.07	0.13	0.29	0.00	0.27	0.00
MASI	0.02	0.03	0.03	0.00	0.04	0.00	0.05	0.05
MIDDAL	0.05	0.04	0.05	0.04	0.07	0.17	0.08	0.00
NAUTSUNDVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NYBERGSUND	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OEVREVATN	0.00	0.07	0.06	0.12	0.12	0.00	0.00	0.00
OEVSTEVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ORSJOREN	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.25
POLMAK	0.12	0.12	0.12	0.13	0.03	0.00	0.04	0.00
REINSNOSVATN	0.00	0.00	0.09	0.12	0.00	0.00	0.00	0.00
ROEYKENES	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROSTEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
STORDALSVATN	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
SVARTAVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TANNSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TARALDSVIKELV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TORA	0.08	0.09	0.08	0.04	0.43	0.44	0.19	0.40
VAEKKAVA	0.05	0.06	0.03	0.03	0.22	0.00	0.06	0.00

Figure 6 | Spring CSI values presented for a lead time of 1 day (left) and a lead time of 9 days (right). The colors indicate CSI range, from white indicating no hits (CSI = 0) to shades of blue up to CSI = 0.5.



	Traw_Praw	Tbma_Praw	Traw_Pbma	Tbma_Pbma	Traw_Praw_Qbma	Tbma_Praw_Qbma	Traw_Pbma_Qbma	Tbma_Pbma_Qbma
AARDAL	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
AUSTENAA	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00
BAGN	0.43	0.43	0.33	0.33	0.43	0.43	0.43	0.43
BJOERNEGAARDSSVINGEN	0.00	0.00	0.12	0.12	0.00	0.00	0.00	0.00
BJOERNSTAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BJORDAL	0.38	0.38	0.30	0.36	0.44	0.38	0.38	0.38
BOEYUMSELV	0.29	0.50	0.44	0.50	0.14	0.25	0.17	0.29
BORGAAI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BREKKE_BRU	0.50	0.33	0.33	0.33	0.25	0.25	0.50	0.00
BULKEN	0.43	0.43	0.43	0.30	0.44	0.44	0.38	0.43
DALSBOEVATN	0.42	0.42	0.40	0.40	0.46	0.46	0.40	0.40
DJUPEVAD	0.00	0.00	0.21	0.20	0.20	0.29	0.18	0.10
DYRDALSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FETVATN	0.00	0.17	0.17	0.33	0.17	0.17	0.00	0.00
FISKUM	0.31	0.31	0.27	0.27	0.36	0.36	0.14	0.14
FOENNERDALSVATN	0.00	0.33	0.00	0.40	0.40	0.40	0.50	0.33
GARHAMMERFOSS	0.20	0.20	0.20	0.20	0.25	0.33	0.00	0.00
GJERSTAD	0.29	0.29	0.25	0.25	0.38	0.38	0.00	0.25
GRYTA	0.25	0.25	0.20	0.20	0.33	0.50	0.00	0.00
HAUGLAND	0.33	0.33	0.20	0.22	0.20	0.20	0.20	0.20
HELLELAND	0.40	0.40	0.33	0.40	0.40	0.40	0.40	0.40
HOEGFOSS	0.12	0.12	0.08	0.08	0.25	0.25	0.00	0.00
HOEGGAAS_BRU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KJOELEMO	0.33	0.33	0.25	0.25	0.40	0.40	0.25	0.25
KRAAKFOSS	0.43	0.43	0.33	0.33	0.40	0.33	0.00	0.00
LILLEELV	0.38	0.38	0.25	0.25	0.38	0.38	0.20	0.25
LISLEFJOEDD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MEVATN	0.38	0.38	0.38	0.38	0.40	0.40	0.40	0.40
MOESKA	0.44	0.44	0.29	0.29	0.44	0.44	0.43	0.43
MYGLEVATN	0.25	0.25	0.27	0.27	0.33	0.25	0.33	0.29
NARSJOE	0.40	0.40	0.40	0.40	0.25	0.25	0.25	0.50
NAUTSUNDVATN	0.00	0.00	0.33	0.33	0.33	0.33	0.33	0.33
NESSEDALSELV	0.00	0.00	0.00	0.17	0.00	0.33	0.25	0.20
OEYE	0.00	0.40	0.20	0.33	0.25	0.40	0.25	0.25
OEYUNGEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
REFSVATN	0.33	0.33	0.17	0.17	0.33	0.50	0.33	0.33
REINSNOSVATN	0.38	0.39	0.38	0.40	0.41	0.44	0.53	0.44
ROEYKENES	0.00	0.00	0.17	0.17	0.00	0.00	0.00	0.00
SAETERNBEKKEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
URULA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	Traw_Praw	Tbma_Praw	Traw_Pbma	Tbma_Pbma	Traw_Praw_Qbma	Tbma_Praw_Qbma	Traw_Pbma_Qbma	Tbma_Pbma_Qbma
AARDAL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AUSTENAA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BAGN	0.20	0.20	0.10	0.10	0.20	0.00	0.00	0.00
BJOERNEGAARDSSVINGEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BJOERNSTAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BJORDAL	0.20	0.20	0.25	0.30	0.25	0.25	0.29	0.14
BOEYUMSELV	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00
BORGAAI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BREKKE_BRU	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
BULKEN	0.00	0.00	0.12	0.09	0.14	0.14	0.00	0.00
DALSBOEVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DJUPEVAD	0.00	0.00	0.07	0.06	0.08	0.00	0.00	0.00
DYRDALSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FETVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FISKUM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FOENNERDALSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GARHAMMERFOSS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJERSTAD	0.12	0.12	0.10	0.10	0.00	0.00	0.00	0.00
GRYTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAUGLAND	0.20	0.20	0.10	0.10	0.00	0.00	0.00	0.00
HELLELAND	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HOEGFOSS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HOEGGAAS_BRU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KJOELEMO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KRAAKFOSS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LILLEELV	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00
LISLEFJOEDD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MEVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MOESKA	0.00	0.00	0.11	0.11	0.00	0.22	0.00	0.00
MYGLEVATN	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.00
NARSJOE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NAUTSUNDVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NESSEDALSELV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OEYE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OEYUNGEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
REFSVATN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
REINSNOSVATN	0.00	0.00	0.00	0.22	0.36	0.21	0.30	0.09
ROEYKENES	0.00	0.00	0.17	0.17	0.00	0.00	0.00	0.00
SAETERNBEKKEN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
URULA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 7 | Autumn CSI values presented for a lead time of 1 day (left) and a lead time of 3 days (right). The colors indicate CSI range, from white indicating no hits (CSI = 0) to shades of blue up to CSI = 0.5.

The effect of pre- and postprocessing for a selection of events and catchments

Well-forecasted streamflow is essential to determine a correct flood warning level. In this subsection, we present three flood events and catchments to demonstrate how the different processing approaches influence the ensemble flood forecasts, and how they correspond to warning levels and the reference streamflow.

Figure 8 shows the outcome of the different processing approaches for the atmospheric river event in the October 2014 event at Bulken (see also Figure 1, Table 1) in western Norway. Some of the ensemble members reach the reference streamflow (black line) when P_{bma} is applied without Q_{bma}. However, none of the ensemble medians reach up to the threshold warning level exceeded by the reference streamflow (black line). For some members, P_{bma} induces very large streamflow forecasts, whereas postprocessing removes the effect of P_{bma} (Figure 8 left and right, respectively).

Figure 9 shows the outcome of the different processing approaches for the extreme weather event *Synne* hitting southern Norway in early December 2015. We see that precipitation is underestimated by the raw forecasts, and none of the PS result in ensemble members that reach the reference level for streamflow. The same pattern is seen for Moeska as for Bulken, where P_{bma} induces high streamflow values (Figure 9 left) that are later suppressed by the Q_{bma} (Figure 9 right).

Figure 10 shows the outcome of the different processing approaches for the snowmelt flood in May 2014 at Nybergsund in eastern Norway. For this flood, there are minimal differences between the PS. However, postprocessing reduces the median forecasts for all lead times, in addition to increasing the spread. In this case, the hydrological model might lack snow and is

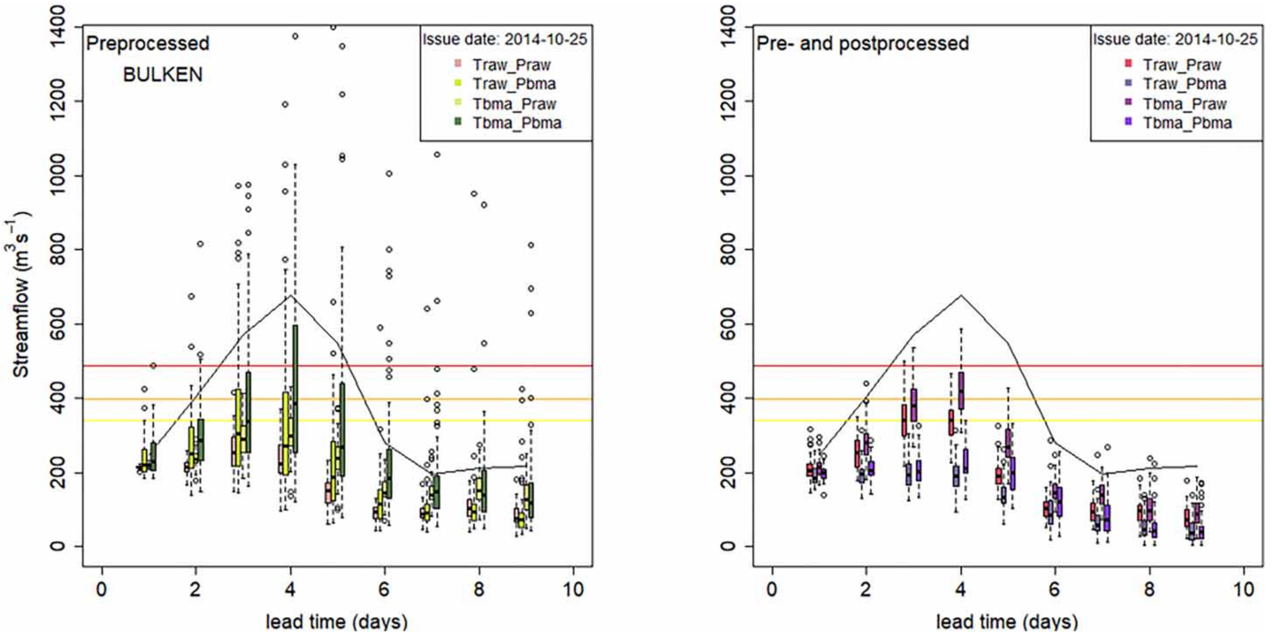


Figure 8 | Bulken catchment: hydrological ensemble forecast of an atmospheric river event 4 days before the peak of the event. The figures present the processing alternatives and the reference streamflow. Colored horizontal lines indicate the operational warning thresholds: yellow (mean annual flood), orange (5-year flood), red (50-year flood), based on the model simulated return levels.

therefore not able to produce snowmelt for streamflow (no effect by T_{bma}), and/or lack of precipitation in the weather forecasts.

Table 5 shows the number of members that for each processing approach exceed the warnings threshold for the events presented in Figures 8–10. Included are the three lead times with the highest warning level for each event.

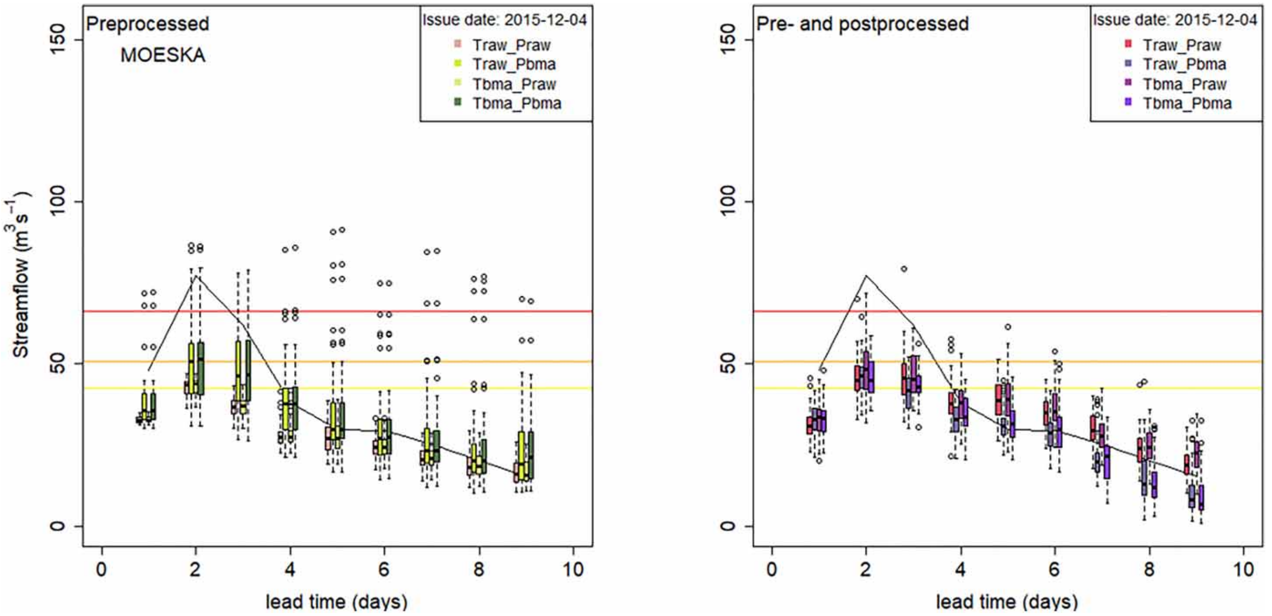


Figure 9 | Moeska catchment: hydrological ensemble forecasts for an extreme weather event 2 days before the peak of the event. The figures present the processing alternatives and the reference streamflow. Colored horizontal lines indicate the operational warning thresholds: yellow (mean annual flood), orange (5-year flood), red (50-year flood), based on the model simulated return levels.

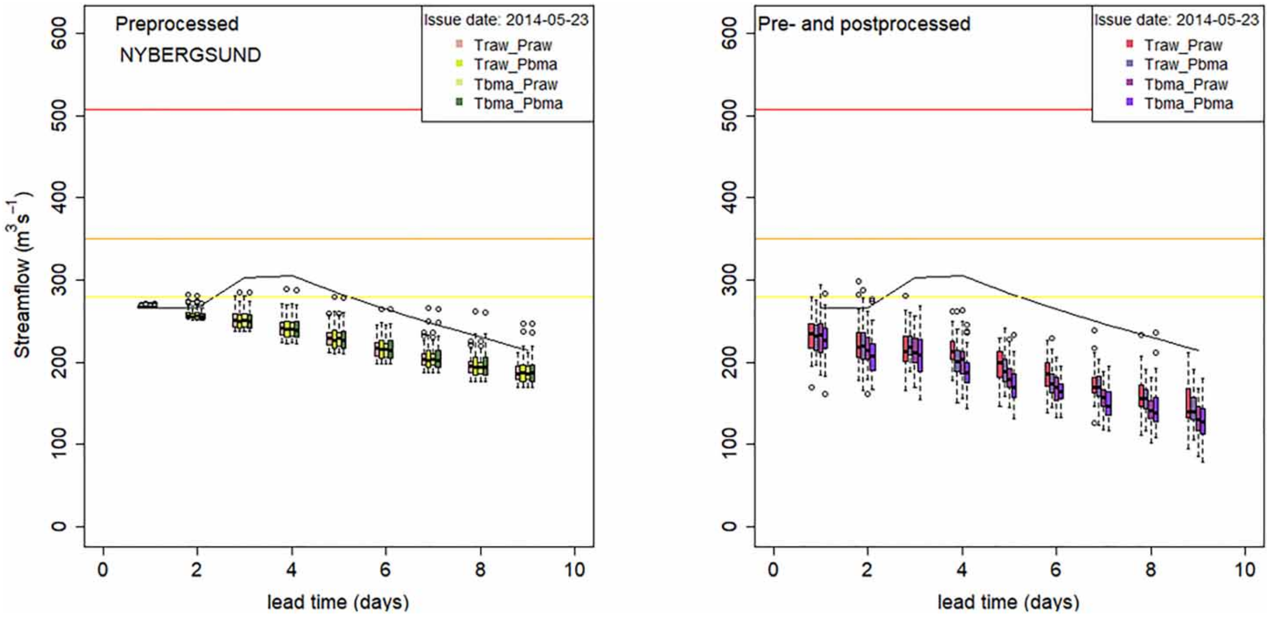
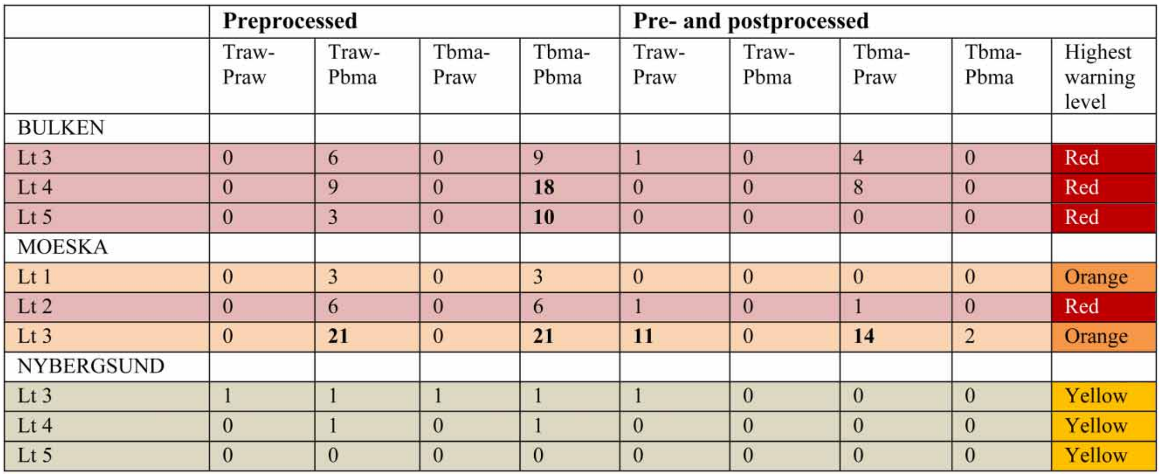


Figure 10 | Nybergsund catchment. Hydrological ensemble forecasts for a snowmelt event four days before the peak of the event. The figures present the processing alternatives and the reference streamflow. Colored horizontal lines indicate the operational warning thresholds: yellow (mean annual flood), orange (5-year flood), red (50-year flood), based on the model simulated return levels.

Table 5 | The number of ensemble members exceeding the highest warning threshold for each of the processing methods for the three flood events shown in Figures 8–10



	Preprocessed				Pre- and postprocessed				
	Traw-Praw	Traw-Pbma	Tbma-Praw	Tbma-Pbma	Traw-Praw	Traw-Pbma	Tbma-Praw	Tbma-Pbma	Highest warning level
BULKEN									
Lt 3	0	6	0	9	1	0	4	0	Red
Lt 4	0	9	0	**18**	0	0	8	0	Red
Lt 5	0	3	0	**10**	0	0	0	0	Red
MOESKA									
Lt 1	0	3	0	3	0	0	0	0	Orange
Lt 2	0	6	0	6	1	0	1	0	Red
Lt 3	0	**21**	0	**21**	**11**	0	**14**	2	Orange
NYBERGSUND									
Lt 3	1	1	1	1	1	0	0	0	Yellow
Lt 4	0	1	0	1	0	0	0	0	Yellow
Lt 5	0	0	0	0	0	0	0	0	Yellow

Three lead times (Lt) are presented for each catchment.

DISCUSSION AND CONCLUSION

The results demonstrate that all catchments benefitted from one or more of the applied PS, thereby confirming our working hypothesis. However, it was not possible to identify a distinct processing chain that is optimal for all forecasts. The optimal method varies with several factors including lead time and season. The flood-generating process is often seasonal, i.e., snowmelt floods are more prone in spring and for inland and high elevation catchments, and rain-induced floods are more typical for autumn and in coastal catchments.

Part of answering our first research question '*How should pre- and postprocessing be combined to improve streamflow forecasts with an emphasis on floods*?' is that postprocessing alone seems to be the least optimal choice when evaluating both the full dataset and even less optimal when the subset of floods is considered. This approach is significantly worse than the best processing approach, both for floods and for all streamflow. This clearly demonstrates the importance of correcting biases and spread in the forcing variables. The catchments' responses to the temperature and precipitation inputs are nonlinear, in particular for snow accumulation and snow melt processes where temperature thresholds are important. Using postprocessing alone is therefore less effective in correcting for biases in inputs to the hydrological model. We find that for the full dataset, the best performance is seen when applying postprocessing combined with preprocessing of temperature for lead times of up to three days, whereas for the longer lead times preprocessing of temperature alone or both precipitation and temperature provide the best performance. Global meteorological ensembles often lack spread for shorter lead times since they are designed for medium-range forecasts and therefore use perturbations that optimize the ensemble spread for longer lead times. BMA models used both for pre- and postprocessing will therefore improve the forecast skill. It would be instructive to assess whether using meteorological ensembles from a regional weather model, which are better able to model the uncertainties in the short range compared to the ensembles from global weather models (Frogner *et al.* 2019a, 2019b), as inputs to the hydrological model alter this finding. However, such forecasts were not available for our study period.

The improvement in skill resulting from the PS is smaller for the flood dataset compared to the complete dataset, and for some catchments, the processing deteriorates the forecasts (Figure 3). We find that postprocessing is less useful for the three first lead times for the flood dataset as compared to the full dataset. Preprocessing both precipitation and temperature for the shortest lead times and only temperature for the longest lead times was the best choice for the largest portion of the catchments in the flood dataset. This result is in line with Benninga *et al.* (2017) who underline the importance of improving the meteorological inputs, in particular for high flow events. In addition to the differences in preferred PS between catchments, we find that for a single catchment, the best processing scheme varies with lead time (i.e., Figures 6 and 7). This underlines that forecast errors arise from different sources, and that being conclusive based on relatively small sample of floods is difficult. The results further showed that autumn floods were particularly difficult to predict beyond a lead time of 3 days, where processing did not improve the flood prediction capability for 28 of 33 catchments with a CSI of zero (Figure 7 right).

Answering our second research question '*Are there regional seasonal patterns in the preferred combination of pre- and postprocessing approaches?*', the results show that the preferred scheme has both regional and seasonal patterns when evaluated for the flood dataset. The regional pattern shows that catchments benefitting from preprocessing alone are, to a large degree, located in coastal areas whereas postprocessing is more important for the inland and high-elevation catchments where temperature and slower snowmelt processes dominate (Figure 4). Furthermore, P_{bma} is the most successful processing scheme in areas with high precipitation (i.e., the west and south-west coast of Norway).

The performance of the PS has clear seasonal patterns. The seasonal effect was evaluated by separating spring floods from autumn floods. The CSI shows that there are large differences in predictability between seasons. For autumn floods there is almost no predictability beyond 3 days, whereas in contrast, spring floods show predictability for up to 9 days. These results indicate that the predictability of floods depends on the flood-generating processes, i.e., snowmelt-induced spring floods are easier to forecast than rain-induced autumn floods. These results further imply that the autumn precipitation and floods are the most difficult to predict and have the highest potential for improvements. Typical catchments improved by BMA applied to precipitation (P_{bma}) are located in coastal and western Norway and are hence prone to high precipitation amounts. One concern when using BMA for preprocessing precipitation is that some of the ensemble members in P_{bma} attained physically nonplausible values, resulting in very high flood forecasts. This is apparent for the Bulken catchment for the October 2014 event (Figure 8). The explanation is that the Bulken catchment experienced large amounts of precipitation during a preceding event. Several of the raw ensemble members for this preceding event had much lower precipitation than what was later observed, whereas the high precipitation for the October 2014 event was better forecasted. Consequently, the BMA procedure increased the forecasted precipitation values too much. In addition, the use of a positively skewed gamma distribution for the kernel amplifies high precipitation values. We believe that this effect can be particularly important in western Norway where small shifts in wind directions might significantly change spatial precipitation patterns and thereby introduce a potential for large errors in forecasts. Possible solutions could be to use a categorical approach (e.g., Ji *et al.* 2019), where the precipitation is separated into precipitation categories (based on for example daily ensemble mean) and unique BMA models are trained for each category.

Cold climate challenges in flood forecasting are demonstrated by the importance of correct temperature and precipitation forecasts for snow storage estimations. For both Bulken and Moeska (Figures 8 and 9) preprocessing temperature affects streamflow through the snowmelt. This indicates that the models have snow available in higher elevated parts of the catchment. On the other hand, neither P_{bma} nor T_{bma} affected the streamflow for the snowmelt flood in Nybergsund. In this example, there is no snow in the model's internal state and therefore, in a situation of snowmelt, any increase in temperature by T_{bma} will not increase streamflow.

For the calculation of CSI, we used a limit of 10 ensemble members (a probability of about 20%) exceeding the flood threshold to issue a flood warning. The ensemble can provide a whole range of probabilities and here we only evaluated for one probability level. The optimal probability of exceedance to issue a flood warning might be different between catchments, lead times, and seasons. Another aspect is to investigate the acceptance level for false alarms to missed events. The number of tolerable false alarms might depend on the impacts of the event (e.g., risk evaluation), and it is therefore difficult to make one absolute decision on behalf of all possible exceedance levels (flood sizes) and affected parties. We acknowledge that the choice of evaluation criteria can be different depending on the users and the cost of mitigation action compared to the loss due to an event, and that false alarms and missed events might be weighted differently depending on a total cost-loss evaluation.

We conclude:

- An evaluation of CRPS for the complete dataset of 2 years showed that the combination of pre- and postprocessing is most effective for short lead times, up to 2–3 days. For longer lead times, PS that only include preprocessing provide the best results, either BMA applied to temperature (T_{bma}) alone or in combination with precipitation (P_{bma}).
- For the flood dataset, the added value of processing is less clear. Overall, the best approach for all lead times is to preprocess both precipitation and temperature.
- The processing is sensitive to regional patterns. Postprocessing was most effective for inland and higher elevated catchments whereas the coastal catchments gained more from preprocessing. BMA applied to precipitation and temperature improved CRPS for the western and southwestern coastal catchments for the early lead times, whereas T_{bma} was most important for the longer lead times.
- We see a substantial difference in performance between spring and autumn floods using critical success index (CSI) for evaluation. In autumn, there is almost no predictive skill for lead times of more than 3 days. Spring floods have a higher predictability for up to 9 days in advance.
- The focus for further improvements should be on the preprocessing of high precipitation rates. For most incidents, the highest precipitation incidents and hence floods were underestimated, whereas for a few incidents, preprocessing high precipitation rates resulted in unrealistic amounts for individual ensemble members.

ACKNOWLEDGEMENTS

The authors would like to thank Bård Grønbech at NVE for help setting up the hydrological model used in this study.

DATA AND SCRIPTS

We made use of the following R-packages: ncdf4, ensembleMOS, ensembleBMA, SpecsVerification.

The code available at https://github.com/metno/fimex was used for the resampling and reprojection of the gridded datasets.

The SeNorge data are downloadable, https://thredds.met.no/thredds/projects/senorge.html, Met Norway.

The ensemble forecast data are available from ECMWF, and streamflow observation is available from NVE upon request.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Bates, B. C. & Campbell, E. P. 2001 A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resources Research* **37** (4), 937–947.

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

Benninga, H. J. F., Booij, M. J., Romanowicz, R. J. & Rientjes, T. H. 2017 Performance of ensemble streamflow forecasts under varied hydrometeorological conditions. *Hydrology and Earth System Sciences* **21** (10), 5273. https://doi.org/10.5194/hess-21-5273-2017.

Bergström, S. 1976 *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*. Swedish Meteorological and Hydrological Institute, Norrköping.

Box, G. E. P. & Cox, D. R. 1964 An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

Buizza, R. 2015 Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Monthly Weather Review* **125** (1), 99–119.

Buizza, R., Milleer, M. & Palmer, T. N. 1999 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.

Cloke, H. L. & Pappenberger, F. 2009 Ensemble forecasting: a review. *Journal of Hydrology* **375** (3), 613–626.

Duan, Q., Ajami, Q. H., Gao, X. & Sorooshian, S. 2007 Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources* **30**, 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014.

ECMWF 2018 *Changes in ECMWF Models*. Available from: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model.

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

Fraley, C., Raftery, A. E. & Gneiting, T. 2010 Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review* **138** (1), 190–202.

Frogner, I. L., Singleton, A. T., Køltzow, M. Ø. & Andrae, U. 2019a Convection-permitting ensembles: challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society* **145** (Suppl. 1), 90–106. https://doi.org/10.1002/qj.3525.

Frogner, I. L., Andrae, U., Bojarova, J., Callado, A., Escribà, P., Feddersen, H., Hally, A., Kauhanen, J., Randriamampianina, R., Singleton, A., Smet, G., van der Veen, S. & Vignes, O. 2019b HarmonEPS – the HARMONIE ensemble prediction system. *Weather and Forecasting* **34**, 1909–1937. https://doi.org/10.1175/WAF-D-19-0030.1.

Gneiting, T., Raftery, A. E., Westveld III., A. H. & Goldman, T. 2005 Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **133** (5), 1098–1118.

Gneiting, T., Balabdaoui, F. & Raftery, A. E. 2007 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.

Gottschalk, L., Jensen, J. L., Lundquist, D., Solantie, R. & Tollan, A. 1979 Hydrologic regions in the Nordic Countries. *Hydrology Research* **10** (5), 273–286.

Gusong, R. 2016 Personal comment 15.06.2016 [Calibration of HBV – NVE flood forecasting].

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

Hanssen-Bauer, I., Førland, E. J., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J. E. Ø., Sandven, S., Sandø, A. B. & Sorteberg, A. 2017 *Climate in Norway 2100 – A Knowledge Base for Climate Adaption. Technical Report 1*. Norwegian Climate Service Centre.

Hegdahl, T. J., Engeland, K., Steinsland, I. & Tallaksen, L. M. 2019 Streamflow forecast sensitivity to air temperature forecast calibration for 139 Norwegian catchments. *Hydrology and Earth System Sciences* **23** (2), 723–739.

Hersbach, H. 2000 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.

Ji, L., Zhi, X., Zhu, S. & Fraedrich, K. 2019 Probabilistic precipitation forecasting over East Asia using Bayesian model averaging. *Weather and Forecasting* **34**, 377–392. https://doi.org/10.1175/WAF-D-18-0093.1.

Jolliffe, I. T. & Stephenson, D. B. 2012 *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Oxford.

Leith, C. E. 1974 Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review* **102** (6), 409–418.

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.

Lorenz, E. N. 1969 The predictability of a flow which possesses many scales of motion. *Tellus* **21** (3), 289–307.

Madadgar, S., Moradkhani, H. & Garen, D. 2014 Towards improved post-processing of hydrologic forecast ensembles. *Hydrological Processes* **28** (1), 104–122.

Mohr, M. 2008 *New Routines for Gridding of Temperature and Precipitation Observations for 'SeNorge. no'. Met. no Report, 8*.

Nash, J. E. & Sutcliffe, J. V. 1970 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.

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S. & Thielen, J. 2017 The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy* **51**, 278–291. doi: 10.1016/j.envsci.2015.04.016.

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

Raftery, A. E., Gneiting, T., Balabdaoui, F. & Polakowski, M. 2005 Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133** (5), 1155–1174.

Roulin, E. & Vannitsem, S. 2015 Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrological Processes* **29** (6), 1434–1449. https://doi.org/10.1002/hyp.10259.

Sælthun, N. R. 1996 *The Nordic HBV Model*. Norwegian Water Resources and Energy Administration Publication, Vol. 7, Oslo, pp. 1–26.

Schefzik, R., Thorarinsdottir, T. L. & Gneiting, T. 2013 Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science* **28**, 616–640. https://doi.org/10.1214/13-STS443.

Seierstad, I., Kristiansen, J. & Nipen, T. 2016 Better temperature forecasts along the Norwegain coast, newsletter, 148. Available from: https://www.ecmwf.int/en/newsletter/148/news/better-temperature-forecasts-along-norwegian-coast (accessed 1 February 2019)

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P. & Mejia, A. 2018 Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. *Hydrology and Earth System Sciences* **22**, 1831–1849. https://doi.org/10.5194/hess-22-1831-2018.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T. & Fraley, C. 2007 Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135** (9), 3209–3220.

Stohl, A., Forster, C. & Sodemann, H. 2008 Remote sources of water vapor forming precipitation on the Norwegian west coast at 60°N–a tale of hurricanes and an atmospheric river. *Journal of Geophysical Research: Atmospheres* **113** (D5), DO5102. https://doi.org/10.1029/2007JD009006.

Thyer, M., Kuczera, G. & Wang, Q. J. 2002 Quantifying parameter uncertainty in stochastic models using the Box–Cox transformation. *Journal of Hydrology* **265** (1–4), 246–257.

Tveito, O. E. 2007 Spatial distribution of winter temperatures in Norway related to topography and large-scale atmospheric circulation. In: *Proceedings of the PUB Kick-off Meeting Held in Brasilia*, 20–22 November 2002. IAHS Publications, Vol. 309, pp. 186–194.

Tveito, O. E., Bjørdal, I., Skjelvåg, A. O. & Aune, B. 2005 A GIS-based agro-ecological decision system based on gridded climatology. *Meteorological Applications.* **12**, 57–68. https://doi.org/10.1017/S1350482705001490.

UNISDR 2004 Guidelines for Reducing Flood Losses, United Nations International Strategy for Disaster Reduction, DRR7639 UNISDR. Available from: http://www.unisdr.org/we/inform/publications/558.

Vannitsem, S., Wilks, D. S. & Messner, J. W. 2018 *Statistical Postprocessing of Ensemble Forecasts*. Elsevier. ISBN 9780128123720, doi: 10.1016/B978-0-12-812372-0.09988-X.

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

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

Wetterhall, F., Pappenberger, F., Alfieri, L., Cloke, H. L., Thielen-del Pozo, J., Balabanova, S., Daňhelka, J., Vogelbacher, A., Salamon, P. & Carrasco, I. 2013 HESS opinions 'Forecaster priorities for improving probabilistic flood forecasts'. *Hydrology and Earth System Sciences* **17** (11), 4389–4399. https://doi.org/10.5194/hess-17-4389-2013.

Xu, J., Anctil, F. & Boucher, M. A. 2019 Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. *Journal of Hydrology* **578**, 124002.

Yang, J., Reichert, P., Abbaspour, K. C. & Yang, H. 2007 Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. *Journal of Hydrology* **340** (3–4), 167–182.

Zalachori, I., Ramos, M. H., Garçon, R., Mathevet, T. & Gailhard, J. 2012 Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science and Research* **8** (1), 135–141. https://doi.org/10.5194/asr-8-135-2012.

Zappa, M., Jaun, S., Germann, U., Walser, A. & Fundel, F. 2011 Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research* **100**, 246–262. doi:10.1016/j.atmosres.2010.12.005.

Zhu, Y. & Newell, R. E. 1998 A proposed algorithm for moisture fluxes from atmospheric rivers. *Monthly Weather Review* **126**, 725–735. https://doi.org/10.1175/1520-0493(1998)126 < 0725:APAFMF > 2.0.CO;2.

First received 11 February 2022; accepted in revised form 17 January 2023. Available online 4 February 2023